UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Doble Grado en Ingeniería Informática y Matemáticas

TRABAJO FIN DE GRADO

PREDICCIÓN DE LA DESAPARICIÓN DE ENLACES EN REDES SOCIALES

Beatriz Romera de Blas
Tutor: Pablo Castells Azpilicueta

JULIO 2017

PREDICCIÓN DE LA DESAPARICIÓN DE ENLACES EN REDES SOCIALES

AUTOR: Beatriz Romera de Blas
TUTOR: Pablo Castells Azpilicueta

Grupo de Recuperación de Información
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio de 2017

Resumen

La llegada del siglo XXI ha traído consigo un cambio radical en la forma en que las personas emplean la tecnología para relacionarse entre ellas. La aparición de las grandes plataformas de redes sociales online como Facebook, Twitter, Instagram o LinkedIn ha supuesto una revolución en este terreno. La posibilidad de obtener gran cantidad de información sobre estas redes sociales y ser capaces de modelar y comprender la evolución de las mismas ha impulsado grandes áreas de investigación en este ámbito. El presente trabajo explora uno de los aspectos aún apenas estudiado en este campo: la predicción o recomendación de enlaces con potencial de desaparición en la red de un usuario.

Este proyecto se centra por un lado en estudiar las métricas propias del análisis de redes sociales sobre los enlaces persistentes y desaparecidos de un grafo y desarrollar recomendadores de ranking fundamentados en esas métricas en base a los resultados de dicho estudio. Otras dos aproximaciones al problema de recomendación son exploradas: a) una inversión de algoritmos clásicos de recomendación de contactos y b) la adaptación al formato de recomendación de métodos supervisados de aprendizaje automático basados en características de los usuarios.

Llevamos a cabo asimismo una evaluación comparativa de estas tres variedades de predictores, en términos de la calidad y relevancia media de las sugerencias ofrecidas a los usuarios, mediante métricas de evaluación tales como Precisión y Recall.

La parte empírica de este trabajo se ha realizado mediante experimentos offline sobre dos muestras extraídas de la red social Twitter: un grafo de contactos explícitos (red de follows) y un grafo de interacciones entre usuarios (retweets, menciones y respuestas), a fin de estudiar, como parte de los objetivos principales de este trabajo, las semejanzas y diferencias que se observan entre estos dos tipos de grafos de distinta naturaleza.

Palabras clave: Evaluación, desaparición, grafo, métricas, recomendación, red social, predicción de enlaces, Twitter.

Abstract

The arrival of the 21st century has brought along a radical change in the way people use technology to interact between each other. The emergence of the big online social networks such as Facebook, Twitter, Instagram or LinkedIn has led to a revolution in this field. The possibility of obtaining big amounts of information about these social networks and being able to model and understand their evolution, has boosted the investigation in that matter. This work explores one of the areas that has not been deeply studied yet: the prediction or recommendation of links with potential of disappearance on the user network.

This project is focused, on one hand, on the study of social network analysis metrics over the persistent and disappeared links of a graph, along with the development of ranking recommendators based on these metrics using the result of that research. Another two approaches to the problems are explored: a) an inversion of the classic contact recommendation algorithms and b) the adaptation to the recommendation format of supervised machine-learning methods based on user characteristics.

We carry out also a comparative evaluation of these three recommendators varieties of predictors, interms of average quality, accuracy and relevancy of the suggestions offered to users, using evaluation metrics such as Precision and Recall.

The empirical part of this work has been performed through offline experiments over two samples obtained from the Twitter social network: one graph of explicit friendship relations (follows network) and one graph of user interactions (retweets, mentions and replies), with the aim of studying, as part of the main goals of these project, the similarities and differences between these two different nature graphs.

Keywords: Evaluation, disappeared, graph, link prediction, metrics, recommendation, social network, Twitter.

Agradecimientos

En primer lugar, me gustaría agradecer a mi tutor, Pablo Castells, la posibilidad de desarrollar este trabajo y por la ayuda recibida estos meses. También quiero dar las gracias a Javi, miembro del Grupo de Recuperación de Información, por su guía y ayuda en los problemas más técnicos.

Agradecer también a mis compañeros de carrera, sin los que estos 5 años no habrían sido lo mismo, y con la ayuda de los cuales he llegado hasta aquí, en especial a Pedro, mi compañero de prácticas durante todos estos años y un gran apoyo en el área matemática.

Por último, dar las gracias a mi familia por su constante apoyo, y por enseñarme que con esfuerzo y dedicación no hay ninguna meta imposible.

ÍNDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

La predicción de enlaces es gran un ámbito de estudio en el campo del análisis de redes, y en particular en el de las redes sociales. Los enlaces de una red no aparecen y desaparecen de forma aleatoria, tal como demuestra Newman (2003) en su estudio sobre la diferencia entre las redes aleatorias y las redes 'del mundo real'. Predecir qué factores o características de la red favorecen la aparición de una conexión entre dos nodos o su desaparición ha sido y sigue siendo objeto de investigación (Wang et al. 2015), y en este contexto se enmarcan los objetivos de este trabajo.

La predicción de enlaces, de forma general, consiste en intentar determinar qué enlaces van a crearse en una red o cuáles de los enlaces ya existentes van a desaparecer en el futuro. Esto es relevante desde el punto de vista de la comprensión de la evolución de las redes, en las cuales hay una dinámica continua de enlaces que aparecen y desaparecen y que modelan dicha evolución.

El problema más abordado hasta el momento en este contexto ha sido el de la aparición de enlaces, que ha visto su principal aplicación en la elaboración de recomendadores de contactos, una herramienta que, en base a la red de amistades de un usuario, o información personal propia, permite al usuario de forma sencilla crear nuevas conexiones con personas que pueden ser de su interés, a través de la sugerencia de las mismas. Las grandes redes sociales como Facebook, Twitter y LinkedIn incorporan estas funcionalidades desde hace tiempo (Gupta et al. 2013). Sin embargo, el estudio de la desaparición de enlaces, aún en una etapa más joven, no parece tener su lugar todavía en el mundo online de la recomendación.

El problema de la detección de enlaces con mayor potencial de desaparición es sin embargo también relevante (Guns 2009), y puede encontrar su aplicación en aspectos tales como ayudar al usuario a depurar su lista de contactos en una red social o a intentar retomar el contacto con alguna persona antes de perder la conexión definitivamente. También puede ser de utilidad para los denominados Social Media Managers o Community Managers con el objetivo detectar qué grupo de seguidores pueden estar perdiendo el interés por su producto y podrían cortar la conexión con ellos. Otro gran campo de aplicación sería a nivel de una empresa detectar qué clientes están planeando dejar de contar con sus servicios y separarse por tanto de la red de dicha empresa, de tal modo que una temprana detección ayudaría a intentar tomar medidas para retenerlos.

Por todos estos motivos y por la limitada investigación que se ha llevado a cabo de momento en este campo, en este trabajo se intenta profundizar más en este problema de detección y recomendación de enlaces que van a desaparecer de una red. El estudio se centra en dos tipos de redes, una red social estable o de conexiones de amistad y en una red más dinámica, una red de interacciones entre usuarios, donde cada interacción representa un enlace entre dichos usuarios. De cara a la vertiente experimental, nuestro estudio se particulariza en la red social de Twitter.

Para el fin planteado, partimos de un análisis de métricas de caracterización de redes sociales sobre los enlaces que desaparecen y persisten entre dos capturas temporales de la red objeto

de estudio, para con esas bases elaborar y comparar recomendadores fundamentados en dichas métricas.

Por otra parte, exploramos la inversión de métodos del estado del arte en recomendación de contactos, partiendo de la hipótesis de que los métodos que funcionan bien para encontrar personas que pueden con alta probabilidad conectarse a un usuario, si son invertidos, detectarán qué personas podrían desconectarse del usuario. El trabajo finaliza con una tercera aproximación al problema de recomendación basada en una adaptación de métodos de aprendizaje automático sobre características personales del usuario y de su actividad en la red social en la que se encuentra presente.

1.2 Objetivos

El objetivo genérico de este trabajo es desarrollar métodos de predicción de la desaparición de conexiones en redes sociales dentro del ámbito de la recomendación.

Los objetivos específicos que se persiguen son:

- Planteamiento de una formulación matemática del problema que englobe también a la predicción de la aparición de conexiones en redes sociales.
- Análisis empírico de las características y métricas que distinguen a las conexiones que desaparecen en una red social de las que permanecen.
- Desarrollo de métodos de predicción basados en rankings por métricas de análisis de redes sociales.
- Estudio de la inversión de métodos de predicción de la aparición de conexiones como potenciales recomendadores de la desaparición de conexiones.
- Adaptación de predicciones obtenidas a partir de métodos de aprendizaje automático al problema de recomendación.
- Analizar las diferencias, en el problema y las soluciones propuestas, entre redes sociales estables y redes de interacción.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- **Capítulo 1 – Introducción.** Se motiva y plantea el problema y se indica la dirección específica a desarrollar.
- **Capítulo 2 – Estado del arte.** Se resume el trabajo relacionado más relevante abarcando las áreas de los sistemas de recomendación, el análisis de redes sociales, los métodos de evaluación asociados y el trabajo similar al desarrollado en este proyecto
- **Capítulo 3 – Formulación.** Se formula el problema y se definen los métodos específicos que serán utilizados para su resolución
- **Capítulo 4 – Experimentos.** Se muestran los resultados de las distintas pruebas llevadas a cabo y las conclusiones obtenidas de cada una de ellas, así como una descripción de los datos y herramientas utilizados para obtenerlas.
- **Capítulo 5 – Conclusiones y trabajo futuro.** Se recogen las conclusiones más relevantes de esta investigación y se indican las futuras vías de desarrollo de la misma.

2 Estado del arte

La predicción de desaparición de enlaces en una red social presentada al usuario en forma de recomendación de potenciales usuarios con los que es posible que pierda la conexión, se encuadra dentro del campo de la recomendación de contactos en redes sociales. Este campo combina el trabajo de los sistemas de recomendación con el análisis de redes sociales.

En este capítulo se ofrece una visión general del trabajo más relevante en estos campos, asociado a los objetivos de este trabajo, así como los distintos enfoques que se le han dado a este problema de predicción.

2.1 Sistemas de recomendación de contactos

Un sistema de recomendación se asemeja a un sistema de recuperación de información (o motor de búsqueda), con la diferencia de que los sistemas de recomendación intentan ofrecer al usuario información que le resulte de utilidad sin que éste haya realizado una consulta. Los sistemas de recomendación empezaron a ser desarrollados a principios de los '90. Originalmente el problema se centró en la recomendación de contenido o artículos, pero la aparición de las redes sociales online dio lugar a este nuevo campo.

Las redes sociales proveen mucha información sobre la estructura que rodea a un usuario que los recomendadores pueden aprovechar para elaborar recomendaciones personalizadas, o no, sobre nuevas conexiones con otros usuarios de la red que podrían ser de interés para el usuario. La mayoría de redes sociales ya dispone de mecanismos de recomendación de contactos, basados tanto en principios de conexiones comunes, como de novedad e información que el usuario proporciona a la plataforma online.

2.1.1 Algoritmos de recomendación

En esta sección se incluyen los algoritmos de recomendación de contactos que se examinarán en este trabajo, tanto desde el punto de vista de su formulación matemática como desde su fundamentación. La función $f_u(v)$ representará la puntuación asignada por cada algoritmo a un usuario *v* concreto cuando se le recomienda al usuario *u*.

Aleatorio

Este algoritmo asigna una puntuación entre 0 y 1 de forma uniforme, independientemente de la información relativa a ambos usuarios, generando por tanto una recomendación completamente aleatoria.

$$f_u(v) = random(0,1)$$

Este recomendador será utilizado para fijar una base sobre los valores mínimos de precisión y eficiencia que un recomendador debe cumplir para considerarse de calidad. También aporta una medida de la dificultad del problema de recomendación sobre una red concreta.

Popularidad

Los recomendadores basados en el concepto de popularidad asignan una puntuación al usuario *v* en función del número de contactos que este posee. En el caso de grafos dirigidos, este número de vecinos de *v*, $\Gamma(v)$, viene dado por el grado incidente del mismo.

$$f_u(v) = |\Gamma_{in}(v)|$$

A pesar de la simplicidad de este recomendador, y su falta de personalización, se ha comprobado que ofrece resultados razonablemente aceptables (Cremonesi et al. 2010).

Adamic-Adar

Esta función de puntuación fue propuesta como una medida de similitud basada en características comunes por Adamic et al. (2003). El algoritmo considera similares a aquellos usuarios que tienen una gran cantidad de contactos en común, dando más importancia a aquellos contactos comunes con un bajo grado, considerando más especial su presencia en ambos círculos de contactos. La adaptación de este algoritmo al problema de recomendación de contactos vendría dada por la función:

$$f_u(v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log|\Gamma(w)|}$$

Vecinos comunes

Estrechamente ligado con la perspectiva dada por Adamic-Adar al problema de recomendación, el método de vecinos comunes, o, también conocido como Friend of a Friend (FOAF) propuesto por Newman et al (2001), utiliza el tamaño del vecindario común de dos usuarios para establecer la función de ranking.

$$f_u(v) = |\Gamma(u) \cap \Gamma(v)|$$

K-vecinos cercanos basado en usuario

Este método colaborativo, conocido también como user-based KNN, está basado en la idea de que usuarios similares tienen gustos o intereses similares. En su adaptación al problema de recomendación, este algoritmo utiliza los *k* vecinos más similares al usuario *u*, al cual se le va a ofrecer la recomendación, que están conectados con el candidato *v* y utilizando la función de similitud establecida entre usuarios se calcula la puntuación para dicho usuario *v*. El conjunto de vecinos del usuario *u* es conocido como su vecindario y se expresa como *N(u)*. La función de ranking de este algoritmo es la siguiente:

$$f_u(v) = \sum_{\substack{w \in N(u) \\ (w,v) \in E}} sim(u, w)$$

Donde una de las aproximaciones más comunes para calcular esa función de similitud es la función de similitud por coseno entre usuarios, que puede ser descrita como:

$$sim(u, v) = \frac{|\Gamma_{out}(u) \cap \Gamma_{out}(v)|}{\sqrt{|\Gamma_{out}(u)||\Gamma_{out}(v)|}}$$

BM25

Este es uno de los modelos probabilísticos más efectivos en el campo de la recuperación de información hasta la fecha. El algoritmo considera que la distribución de frecuencia de los términos en un documento sigue una distribución de Poisson (Sparck Jones et al., 2000). Su adaptación al problema de recomendación considerada para el caso límite cuando $k \to \infty$, denominada como versión extrema de BM25 es la siguiente:

$$f_u(v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{RSJ'(w)}{1 - b + b \frac{|v|}{avg_{v'}(|v'|)}}$$

Donde $b \in [0,1]$ es un parámetro que en su planteamiento original sirve para mitigar el efecto de la longitud de los documentos y RSJ' es la ecuación de Roberston-Spark Jones derivada, que en el problema de recomendación de redes sociales tiene la siguiente forma:

$$RSJ'(w) = \log\left(\frac{|V| - |\Gamma(w)| + 0.5}{|\Gamma(w)| + 0.5}\right)$$

2.2 Análisis de redes sociales

Una red social es un conjunto de personas con una serie de contactos o interacciones establecidas entre ellas (Newman 2003). Estas redes han sido objeto de estudio en diferentes campos como la sociología, la psicología, la biología o la estadística. Los primeros trabajos que ser recogen con una mención explícita a las redes sociales se ubican en el ámbito de las ciencias sociales y datan de finales del siglo XIX (Tönnies 1887, Durkheim 1893).

Sin embargo, fue a principios de los 2000s cuando con la aparición de las plataformas sociales online, las redes sociales tomaron un nuevo significado y abrieron un nuevo horizonte de investigación y análisis en ese campo. Algunas de las plataformas más populares como Facebook, Twitter, LinkedIn o Instagram son usadas todos los días por cientos de millones de personas alrededor de todo el mundo. La posibilidad de acceder a gran parte de la información sobre los usuarios presentes en esas redes y la forma en que crean relaciones o interactúan entre ellos ha supuesto una total revolución en el estudio de las relaciones interpersonales y ha impulsado el estudio y explotación de estos datos por parte de empresas e investigadores.

Algunas de las principales aplicaciones que tiene el análisis de redes sociales son entender cómo se crean y destruyen relaciones (Garimella et al. 2014), identificar a personas clave o especiales en una red (Backstrom et al. 2014), detectar comunidades de usuarios, predecir dinámicas sociales, la evolución de una red, cómo frenar una epidemia, maximizar la difusión de un mensaje o elaborar campañas de marketing enfocadas a un público concreto.

Dependiendo de la estructura que tienen las conexiones que los usuarios pueden establecer entre ellos en una determinada red social, podemos hablar de dos tipos de redes:

- **Redes dirigidas o asimétricas:** En estas redes las conexione que se establecen entre los individuos no tienen por qué ser recíprocas. Ejemplos de este tipo de redes son Twitter o Instagram, donde las relaciones entre usuarios son representadas por medio de enlaces dirigidos en el grafo que modela dicha red.

- **Redes no dirigidas o simétricas:** Estas redes se caracterizan porque todas las conexiones que se establecen entre los individuos son recíprocas. Ejemplos de estas redes son Facebook o LinkedIn, donde las relaciones de amistad entre los usuarios son representadas mediante enlaces no dirigidos en el grafo asociado.

2.2.1 Métricas

Para entender la topología o estructura de estas redes sociales existen una serie de métricas que ayudan a caracterizarlas en su conjunto o que estudian las propiedades de sus nodos y enlaces. En esta sección se enuncian algunas de estas características, así como la importancia y relevancia que dichas propiedades pueden tener en la red, para establecer por ejemplo usuarios o enlaces clave.

2.2.1.1 Propiedades de los nodos

Grado

El grado de un nodo es el número de enlaces en los que participa. En redes dirigidas el grado de un nodo puede estudiarse como el número de enlaces incidentes en él (*indegree*) o enlaces salientes del mismo (*outdegree*).

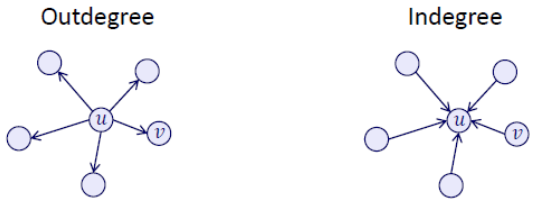

Figura 1. Representación del grado saliente y entrante de un nodo *u*.

A pesar de ser una de las métricas más sencillas, no por ello es menos significativa, ya que a partir del grado de un nodo podemos, por ejemplo, obtener su influencia en la red basada en su popularidad. El estudio de los grados de los nodos de un grafo suele hacerse a través de su distribución, es decir, de la cantidad de nodos que tienen un grado determinado, ofreciendo de esta forma una visión global de cómo están repartidos los grados en el grafo.

Closeness

El closeness de un nodo refleja su posición de influencia en la red por la rapidez para llegar a los demás nodos, por ejemplo, para transmitir información. Un nodo con un alto valor de esta métrica indica que dicho nodo no necesariamente tiene muchos contactos, ni es un punto clave de paso, si no que se encuentra en una posición cercana en promedio a todos los nodos.

Se utilizan diferentes variantes, con ligeras diferencias, una de las más comunes es la que identifica el closeness de un nodo como la inversa de la distancia mínima media:

$$C(u) = \frac{n-1}{\sum_{v \in V} d(u,v)}$$

siendo *d(u,v)* la distancia mínima entre los nodos *u* y *v*.

Cuando una red no es fuertemente conexa, es decir, no se puede acceder a todos los nodos desde cualquiera de los nodos del grafo, todos los usuarios u tendrían $d(u,v) = \infty$ para algún v, y por tanto $C(u) = 0$. Para evitar este problema se suele considerar solo los v en la misma componente conexa y promediar conjuntamente. En este caso n sería el tamaño de la componente conexa en la que se encuentra u.

Betweenness

Esta métrica identifica los nodos que se consideran puntos de paso entre muchos pares de nodos, lo que no implica directamente que tengan muchos contactos. Los nodos con un alto valor de esta métrica tienen una posición de influencia por su papel en el paso de información y su eliminación de la red tiende a crear disrupción del flujo de información. Habitualmente se suele calcular como el ratio promedio de caminos de distancia mínima (CDM) de la red que pasan por un nodo. Para grafos no dirigidos, en su versión sin normalizar, la fórmula usual es:

$$B(u) = \sum_{\substack{v,w \neq u \\ v<w}} \frac{ns_{v,w}(u)}{ns_{v,w}}$$

$$ns_{v,w} \equiv nº\ de\ CDM\ entre\ los\ nodos\ v\ y\ w$$
$$ns_{v,w}(u) \equiv nº\ de\ CDM\ entre\ los\ nodos\ v\ y\ w\ que\ pasan\ por\ u$$

En grafos dirigidos eliminaríamos simplemente la condición $v<w$.
Para normalizar el betweenness basta con dividir el valor del sumatorio anterior entre el número de caminos de distancia mínima existentes en el grafo.

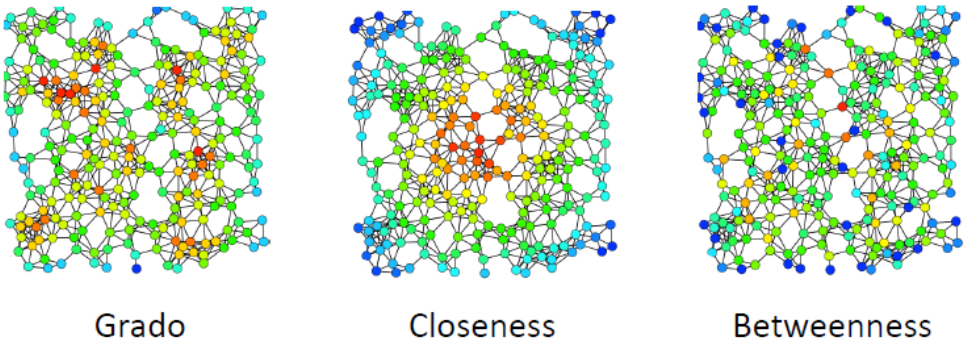

Figura 2. Propiedades de los nodos en un grafo, en escala de colores.[1]

Como se observa en la Figura 2, aunque los conceptos que reflejan estas métricas parecen similares, nodos con un alto grado o un alto closeness no implican un alto valor de betweenness y viceversa.

Coeficiente de clustering local

Esta propiedad refleja la cohesión del entorno de un nodo. Basada en la noción de cierre tríadico, de forma intuitiva intenta captar en qué medida los vecinos de un nodo están conectados entre sí, es decir, como de completo es el grafo entorno al nodo.

[1] Fuente: http://www.martingrandjean.ch/gephi-introduction/ (Accedida 25/06/2017)

Su valor puede ser calculado como la probabilidad de que dos vecinos de un nodo dado tomados al azar sean vecinos:

$$C(u) = p(v \to w | u \to v, u \to w) = \frac{|conexiones\ entre\ vecinos\ de\ u|}{|conexiones\ posible\ entre\ vecinos\ de\ u|} \in [0,1]$$

donde el número de conexiones posibles entre vecinos de *u* se puede calcular como:

$$|conexiones\ posible\ entre\ vecinos\ de\ u| = \begin{cases} g(u)(g(u)-1), & si\ G\ dirigido \\ g(u)(g(u)-1)/2, & si\ G\ no\ dirigido \end{cases}$$

El coeficiente de clustering suele estar inversamente relacionado con betweenness, un alto valor del coeficiente de clustering implica redundancia en la comunicación y por tanto bajo valor de betweenness, mientras que un bajo clustering representa una posición ventajosa en la transmisión de información y por tanto alto betweenness.

PageRank

Marca registrada y patentada por Google que recoge una familia de algoritmos utilizados originalmente para asignar la relevancia de los documentos indexados por un motor de búsqueda. Desarrollado por los fundadores de Google (Brin y Page 1998), este algoritmo, que se actualiza constantemente, sigue formando parte del buscador de Google[2] para determinar la importancia de una página.

La fórmula original (Brin y Page 1998), adaptada a la notación de grafo que se utiliza en este trabajo, es la siguiente:

$$P(v) = \frac{r}{|V|} + (1-r)\sum_{u \to v} \frac{P(u)}{g_{out}(u)}$$

$$r \in (0,1)$$

Donde *r* representa la probabilidad de que un usuario salte a otro aleatorio sin utilizar alguno de sus enlaces disponibles.

Este algoritmo, originalmente ideado para el problema de recuperación de información en motores de búsqueda, se utiliza también en redes sociales como un indicio de la importancia de un usuario, donde un usuario es más importante si tiene muchos enlaces entrantes que provienen de nodos con muy pocos enlaces salientes.

Esta métrica se puede clasificar como una métrica asociada al proceso estocástico que corresponde a las cadenas de Markov de caminos aleatorios (*random walk*), en la que la probabilidad de estar en un determinado estado no depende del estado anterior y la probabilidad de pasar de un estado a otro se puede calcular. El PageRank por tanto de un nodo representaría la probabilidad de encontrarse en un instante dado en dicho nodo.

[2] Motor de búsqueda de Google: www.google.es (Accedida 25/06/2017)

2.2.1.2 Propiedades de los enlaces

Miden el papel de un enlace entre dos usuarios y el efecto que tiene para dichos usuarios y para la red. Gran parte de las medidas giran en torno a la idea de enlace débil y enlace fuerte, donde un enlace débil sería aquel que conecta dos comunidades distintas de usuarios.

Arraigo

También denominado embeddedness de un enlace. Enlaces muy arraigados conectan nodos con muchos enlaces en común. La fórmula para calcular el arraigo de un enlace viene dada por:

$$Arraigo(u,v) = Jaccard(vecinos(u) - \{v\}, vecinos(v) - \{v\}) = \frac{|(vecinos(u) - \{v\}) \cap (vecinos(v) - \{u\})|}{|(vecinos(u) - \{v\}) \cup (vecinos(v) - \{u\})|}$$

Aquellos enlaces con arraigo 0 se les denomina puentes locales, es decir, enlaces que no forman parte de ningún triángulo. Un enlace global sería aquel que si se elimina dividiría el grafo en dos componentes conexas. El arraigo ofrece un punto de vista complementario al coeficiente de clustering local, de forma que los usuarios con muchos enlaces débiles (poco arraigados) tienden a tener un valor bajo de clustering.

Betweenness

Se define de forma similar al betweenness de los nodos. En este caso mide el ratio de caminos de distancia mínima (CDM) de la red que pasan por un enlace, cuya fórmula asociada para grafos no dirigidos, en su versión sin normalizar, es:

$$B(u_1,u_2) = \sum_{\substack{v \neq u_1, w \neq u_2 \\ v<w}} \frac{ns_{v,w}(u_1,u_2)}{ns_{v,w}}$$

$$ns_{v,w} \equiv nº\ de\ CDM\ entre\ los\ nodos\ v\ y\ w$$
$$ns_{v,w}(u_1,u_2) \equiv nº\ de\ CDM\ entre\ los\ nodos\ v\ y\ w\ que\ pasan\ por\ el\ enlace\ entre\ u_1\ y\ u_2$$

En grafos dirigidos eliminaríamos simplemente la condición $v<w$ al igual que en el betweenness de nodos. Y de nuevo la forma de normalizar el valor de esta métrica sería promediarla entre el número de caminos de distancia mínima existentes

Reciprocidad

En un grafo dirigido, un enlace se considera recíproco si dado el enlace *(u,v)*, existe en el grafo el enlace *(v,u).* La existencia de enlaces recíprocos en grafos asimétricos refuerza las conexiones establecidas entre usuarios, permitiendo que la información entre dos usuarios fluya en ambos sentidos. Si todos los enlaces de un grado dirigido fuesen recíprocos el grafo se podría considerar como un grafo no dirigido.

2.3 Técnicas de aprendizaje automático supervisado

Las técnicas de aprendizaje automático supervisado tienen por finalizar deducir una función a partir de datos de entrenamiento. Dichos datos suelen consistir en vectores que contienen distintos campos de entrada y otros con los resultados esperados. La salida de la función puede estar expresada de forma numérica mediante un valor o mediante una etiqueta de clase (en el caso de los clasificadores).

El objetivo de los modelos de aprendizaje automático es crear una función capaz de predecir el valor correspondiente a cualquier entrada, denominada conjunto de test, con el mismo tipo de características que las utilizadas en su entrenamiento, después de haber analizado una serie de ejemplos donde el resultado se encuentra descrito junto a las características propias de entrenamiento.

Esta posibilidad de clasificar ciertos datos a partir de conjuntos previamente estudiados puede resultar de gran utilidad para el problema de predicción de enlaces (Hasan et al. 2006, Lichtenwalter et al. 2010), ya que a través del comportamiento de la red observado con anterioridad se puede predecir la naturaleza futura de los enlaces.

A continuación, se describen brevemente los métodos explorados en este trabajo junto a su idea intuitiva:

Naive Bayes

Algoritmo de clasificación probabilístico basado en el modelo probabilístico bayesiano y que incorpora fuertes suposiciones de independencia. Dada la fórmula de la probabilidad condicionada de Bayes:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Este clasificador asume que las variables condicionan de forma independiente la clasificación de un objeto en una clase u otra, y combina este aspecto con una regla de decisión basada en el máximo a posteriori o MAP.

La fórmula que utiliza el clasificador vendría dada, siendo $f_1, \dots, f_n$ las características estudiadas y *c* la clase que se le asigna a un cierto objeto, por la siguiente expresión:

$$classify(f_1, \dots, f_n) = arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

Regresión logística

Es un tipo de clasificador que utilizando el conjunto de características propias del conjunto de entrenamiento crea un hiperplano definido por una función que divide el espacio entre las distintas clases existentes. Para ello se asigna distintos pesos a las características de forma que se genera una función logística que sirve de umbral para asignar una clase u otra a los objetos a clasificar.

Bosques aleatorios

También conocido como "random forest", es una combinación de árboles de decisión tal que cada árbol depende de los valores de un vector de datos del conjunto de entrenamiento probado de forma independiente. La idea de este clasificador es generar un número determinado de árboles de decisión independientes, elaborados a partir de una selección aleatoria de las características de los vectores y establecer de esa forma una partición entre las dos clases basada en las características, utilizando las clasificaciones realizadas por todos los árboles.

2.4 Evaluación

La investigación y elaboración de sistemas de recomendación va naturalmente ligada con el desarrollo de técnicas que permitan comprobar la efectividad y utilidad de dichas recomendaciones. La búsqueda de métodos de evaluación ha sido una pieza calve en el desarrollo de este campo (Shani et al. 2015) durante décadas, y a día de hoy sigue siendo un campo abierto y de gran interés tanto para investigadores como para los integrantes del área comercial.

A la hora de elegir un método de evaluación es necesario saber qué caracteriza a un algoritmo como mejor que otro, en otras palabras, descubrir qué consideran los usuarios como información de utilidad. Tradicionalmente la habilidad de un algoritmo para ofrecer predicciones más precisas ha sido considerada como la propiedad que define a un buen recomendador (Shani et al. 2015). Sin embargo, hoy en día en el campo de la recomendación hay otras características, como la novedad o la diversidad, que se consideran piezas fundamentales para ofrecer al usuario una recomendación completa (Castells et al. 2015) y evitar elaborar solo recomendaciones dentro de la burbuja de afinidades del usuario, ayudándole a explorar opciones diferentes.

En el caso de la recomendación de eliminación de contactos o detección de potenciales contactos a perder, las técnicas de evaluación más apropiadas están centradas en la visión tradicional previamente mencionada de la precisión, ya que las recomendaciones se realizan sobre elementos pertenecientes al conjunto de conexiones de los usuarios, y por tanto las métricas de novedad y diversidad carecen de sentido. Sin embargo, aún no han sido exploradas técnicas de evaluación específicas para este problema.

A continuación se definen las dos métricas del ámbito de la precisión más populares. Estas dos técnicas están basadas en el concepto de relevancia. En el ámbito de la recomendación de contactos, y en concreto en el campo de estudio de este trabajo, un usuario recomendado *v* se considera relevante para el usuario *u,* al cual se le ha hecho la recomendación, si el usuario *u* elimina la conexión que tenía con el usuario *v* en un tiempo futuro.

Precisión

Esta métrica calcula la fracción de elementos recomendados que es relevante para el usuario objetivo (Baeza-Yates et al. 2011) y está definida como:

$$P(u) = \frac{|Relevant(u) \cap Recommended(u)|}{|Recommended(u)|}$$

Para establecer el número de elementos recomendados que se consideran se utiliza otra versión más común de esta métrica, en la cual se evalúa la precisión de un recomendador para los primeros *k* elementos devueltos:

$$P@k(u) = \frac{|Relevant(u) \cap \{v_1, \dots, v_k\}|}{k}$$

donde $v_1, \dots v_k$ son las *k* primeras recomendaciones del ranking para el usuario *u*.

Este cálculo se promedia entre las recomendaciones generadas para todos los usuarios, *V*, obteniendo un valor medio de precisión para un recomendador, cuyo valor es calculado como:

$$P@k = \frac{1}{|V|} \sum_{u \in V} P@k(u)$$

Recall

Recall computa la proporción de elementos relevantes para el usuario objetivo que ha sido devuelta en una cierta recomendación (Baeza-Yates et al. 2011). Esta métrica puede alcanzar su máximo valor incluso cuando hay elementos irrelevantes en el ranking, su fórmula es la siguiente:

$$R(u) = \frac{|Relevant(u) \cap Recommended(u)|}{|Relevant(u)|}$$

De forma similar a la precisión, existe una versión que permite calcular el valor de esta métrica para los *k* primeros elementos del ranking:

$$R@k(u) = \frac{|Relevant(u) \cap \{v_1, \dots, v_k\}|}{|Relevant(u)|}$$

El valor se promedia entre todos los usuarios para establecer el recall de un determinado recomendador, el cual vendría dado por:

$$R@k = \frac{1}{|V|} \sum_{u \in V} R@k(u)$$

2.5 Trabajo relacionado

El problema de predicción de enlaces que desaparecerán de una red ha sido señalado como un área de investigación de gran interés e importancia (Guns 2009), por dicho motivos algunos investigadores han intentado abordar este problema, objeto de estudio de este trabajo, con distintos enfoques. Esta sección recoge algunas de las principales áreas que se han investigado, así como las distintas perspectivas del problema que han resultado interesantes por el momento.

Guimera et al. (2009) definen el problema como detección de enlaces falsos (*spurious links*), considerando la existencia de un grafo real y un grafo que contiene enlaces no presentes en el grafo real, centrando su investigación en la detección de esos enlaces no fiables. Un trabajo

similar fue llevado a cabo por Pan et al. (2016), en el cual trataban de detectar esos enlaces falsos junto a enlaces previamente desaparecidos en la red. Por su parte, Zeng et al. (2012) intentaron detectar y eliminar esos enlaces falsos procurando que la red en la que se encontraban siguiera conservando sus principales características.

Kivran-Swaine et al. (2011) y Xu et al. (2013) estudiaron el impacto que tiene la estructura de la red de Twitter y las métricas asociadas a sus enlaces para predecir dinámicas de *unfollows* en Twitter, una de las partes que comprende también el presente TFG. Mientras que Kwak et al. (2011) centraron ese análisis en otras características particulares de la red de Twitter, como puede ser la frecuencia con la que un usuario publica tweets o la duración de la relación de seguimiento.

Centrado también en Twitter, pero orientado a la red implícita de interacciones entre los usuarios de dicha plataforma online, Macek et al. (2014) llevaron a cabo un estudio sobre la estabilidad de las interacciones entre usuarios, centrado en observar cómo la cantidad de interacciones iba decayendo con el tiempo entre pares determinados de usuarios.

Estudios con objetivos similares se han llevado a cabo con muestras de otras redes sociales, como por ejemplo Facebook. Es el caso de Quercia et al. (2012), que investigaron qué factores, tanto personales como de estructura de la red, favorecen la pérdida de amistades en dicha red social.

Leskovec et al. (2010) han centrado su investigación en predecir la naturaleza negativa o positiva de los enlaces de una red. Entendemos por enlace negativo aquel en el que un usuario ha mostrado su desagrado o desaprobación hacia otro usuario o algún tipo de contenido producido por dicho usuario. La mayoría de redes sociales solo consideran relaciones positivas, ejemplo de ellas serían las amistades, opción de "me gusta" (Facebook), favorito (Twitter), etc. Sin embargo, el problema de detección de enlaces negativos en redes como YouTube (que permite a los usuarios valorar negativamente un video publicado por otro usuario) podría tener como beneficio añadido detectar la futura desaparición de la conexión entre los dos usuarios involucrados en dicho enlace negativo.

Por otra parte, algunos trabajos como el de Verbeke et al. (2014) han intentado encontrar una aplicación fuera de las redes sociales al problema de detección de enlaces con potencial de desaparición, estudiando el caso de la predicción de pérdida de clientes por parte de empresas.

Como puede observarse, se han llevado a cabo diversas investigaciones de naturaleza similar en los últimos años sobre el tema que aborda este trabajo, sin embargo, aún no se han obtenido conclusiones claras sobre el mismo. Es por este motivo, por el que este trabajo pretende aportar su contribución a este campo mediante la formulación del problema de predicción de la desaparición de enlaces a través de sistemas de recomendación orientados a indicar al usuario cuales de sus enlaces tienen mayor posibilidad de desaparecer en el futuro.

3 Formulación

3.1 Formalización del problema

Sea un grafo dirigido $G = <V, E>$, compuesto por un conjunto de vértices V y los enlaces E existentes entre dichos vértices. Los enlaces se representan mediante el par de vértices que conecta cada enlace, por ejemplo, un enlace e se representa como: $e = (u, v)$, donde u es el vértice origen del enlace e que incide en el vértice destino v.

El problema de la predicción de la aparición o desaparición de enlaces en un grafo plantea estimar mediante distintos métodos la probabilidad de que un enlace nuevo sea creado en el grafo, o que uno de los enlaces existentes sea eliminado del grafo, respectivamente. En el primer caso se consideran todos los enlaces que conectan los vértices del grafo pero no pertenecen al mismo, E^c. En el segundo caso se calculan las probabilidades de desaparición de los enlaces presentes en el grafo, E.

En una red social, y en concreto en el caso de Twitter con el que se realizarán las pruebas, los usuarios serían los vértices que conforman el grafo cuyo conjunto de enlaces correspondería a las conexiones establecidas entre dichos usuarios. Dichos enlaces pueden hacer referencia al establecimiento de una relación de seguimiento por parte de un usuario *u* a un usuario *v*, lo que comúnmente se denomina como una relación de *follow*, este sería el caso de lo que denominaremos **red social estable**. Por otro lado, un enlace también puede representar la interacción que el usuario *u* lleva a cabo con el usuario *v*, ya sea en forma de retweet, mención o contestación (*reply*), en este caso estaríamos ante una **red social de interacciones**.

Por tanto, el problema que aborda este TFG es predecir que conexiones o enlaces entre los usuarios de una red social dada van a dejar de existir en el futuro, es decir, qué usuarios va a dejar de seguir un usuario concreto o con qué usuarios va a dejar de interaccionar

3.2 Métodos

El formato en el que se presentará la información sobre las posibles conexiones que van a desaparecer en la red social es mediante **métodos de recomendación basados en ranking**. Para cada usuario *u* se elabora una recomendación con el ranking de usuarios con los que es más probable que deje de tener una conexión, ordenados por valor decreciente de la puntuación asociada al enlace dirigido entre *u* y el usuario en cuestión, dicho valor corresponde a la salida de un algoritmo de recomendación evaluado para el enlace pertinente.

Respecto a los métodos de **análisis de redes sociales** se estudiarán los valores de las métricas de grado (tanto entrante como saliente), betweenness, coeficiente de clustering y PageRank de los nodos y sobre los enlaces se ha trabajado sobre el arraigo, el betweenness y la reciprocidad de los mismos. Analizando los valores que toman estas métricas para los enlaces y nodos involucrados en enlaces desaparecidos y persistentes. Los detalles sobre la formulación y relevancia de cada una de estas métricas se pueden encontrar en la sección 2.2.1.

Se explorarán también los resultados empíricos de los **algoritmos clásicos** del problema de predicción o **recomendación de contactos** tanto en su versión clásica como en una versión

adaptada al problema que se estudia en este trabajo. Los algoritmos seleccionados para su estudio por su simplicidad y representatividad son los métodos de popularidad, Adamic-Adar, recomendación basada en vecindarios, amigos comunes y BM25 en su versión extrema.

En cuanto a la pequeña aproximación al problema utilizando métodos de **aprendizaje automático supervisado**, se llevarán a cabo pruebas con los algoritmos de Naive-Bayes, regresión logística y bosques aleatorios, transformando las salidas propias de una clasificación mediante aprendizaje automático al formato de recomendación basada en ranking.

Todos estos métodos de recomendación serán evaluados mediante las métricas de precisión y recall, propias del campo de la **evaluación de sistemas de recomendación**.

4 Experimentos

Esta sección contiene las distintas aproximaciones que se ha dado al problema de recomendación de desaparición de contactos, el conjunto de datos y las herramientas con las que se ha trabajado, así como un análisis exhaustivo de los resultados obtenidos en las distintas pruebas realizadas.

4.1 Diseño experimental

4.1.1 Conjuntos de datos

El trabajo empírico de este trabajo se ha centrado en información proveniente de la red social Twitter[3], una plataforma online de microblogging, lanzada en 2006, que permite a los usuarios publicar mensajes (conocidos como *tweets*) de hasta 140 caracteres. Estos mensajes pueden contener, texto, imágenes, URLs o vídeos y pueden ser categorizados utilizando etiquetas denominadas *hashtags*.

Cuando un usuario *u* decide seguir a un usuario *v*, se establece entre ellos una conexión (*u*,*v*), en la cual el usuario *u* es conocido como *follower* y el usuario *v* como *followee*, y la acción de establecer ese enlace entre ellos es denominada *follow*. Es por tanto que decimos que los usuarios de Twitter y las conexiones que establecen entre ellos forman una red social asimétrica, en la que todos los enlaces son dirigidos, es decir, existe un nodo origen y un nodo destino de los mismos.

Cuando un usuario publica un mensaje, éste puede ser visto por todos sus seguidores (*followers*) en sus respectivos muros (*timelines*), esto permite a los usuarios interaccionar entre ellos, por medio de *retweets*, *menciones* o *respuestas*. Un retweet es una forma de compartir con tus seguidores un mensaje creado por otro usuario, una mención (*mention*) consiste en nombrar a alguien por medio de su nombre de usuario al escribir un tweet, de forma que dicho usuario es notificado directamente de que ha sido nombrado y por último una respuesta (*reply*), como su propio nombre indica, permite a los usuarios contestar o responder a los tweets creados por otros usuarios. Por tanto, diremos que existe una conexión entre los usuarios *u* y *v* si *u* ha interaccionado con *v*, es decir, si *u* ha retweeteado, mencionado o respondido a *v*.

A partir de la red de amistades o seguimientos establecida en Twitter se pude obtener un grafo explícito, mientras que a partir de las interacciones llevadas a cabo entre sus usuarios es posible elaborar un grafo implícito, y por consiguiente nuestro estudio experimental se centrará en estos dos grafos. Estudiaremos e intentaremos predecir las dinámicas de desaparición de enlaces en una red social estable (red de follows) y en una red social dinámica (grafo de interacciones).

La gran dimensión de la red completa de Twitter (175 millones de usuarios activos y aproximadamente veinte billones de enlaces en 2012, de acuerdo con Myers et al. 2014), hace imposible para la gran mayoría de personas trabajar con la red social completa. Además, al menos hasta la fecha, solo Twitter tiene acceso al conjunto total de usuarios de la misma.

[3] Twitter: https://twitter.com (Accedida 25/06/2017)

Por este motivo, investigadores y la mayoría de compañías trabajan generalmente con una pequeña muestra de la red manejable computablemente, como ocurrirá en este trabajo de investigación.

4.1.2 Obtención y caracterización del conjunto de datos

La muestra de Twitter con la que se trabajará fue obtenida por el Grupo Recuperación de Información a través de la REST API[4] de Twitter. Para el grafo de interacciones se recogió información en un periodo de tiempo de un mes, entre el 16 de junio de 2015 y el 16 de julio de 2015, a través de una variante del método de muestreo por bola de nieve (Goodman 1961), en la cual empezando por un usuario que sirve como semilla (en este caso @j_yubero) se va expandiendo la colección de usuarios e interacciones recuperadas ampliando el grafo seleccionando un número fijo de tweets publicados más recientemente por el usuario (por ejemplo, el número máximo de tweets que pueden ser recuperados en una simple llamada a la API) o tomando todos los tweets producidos en un intervalo dado de tiempo. Junto con todos estos, tweets se recuperan las interacciones en forma de retweet, mención o respuesta que éstos han generado y los usuarios involucrados en ellas, aumentando así progresivamente el grafo hasta alcanzar el límite de usuarios deseados.

El conjunto total recuperado contenía 10019 usuarios y 2369596 tweets, conjunto que fue dividido en dos particiones temporales (t_0 y t_1) separadas por el final de la tercera semana de recogida de datos (9 de julio de 2015). El grafo de follows, que corresponde a los usuarios en el conjunto previamente mencionado fue obtenido el 9 de octubre de 2015 (t_0) y el 9 de febrero de 2016 (t_1).

La obtención de dos capturas temporales de cada grafo de debe a la necesidad de poder establecer unos conjuntos de entrenamiento y test ya que todas las pruebas que se llevan a cabo durante este trabajo se producen de forma offline.

El grafo de follows consta de 630504 enlaces en t_0, de los cuales 35600 desaparecen, es decir, un 5.65% de los enlaces existentes en t_0 desaparecen en t_1, persistiendo el resto.

El grafo de interacciones consta de 170425 enlaces en t_0, de los cuales 77142 desaparecen, lo que constituye un 45.26% de los enlaces presentes.

Por otro lado, para poder aplicar técnicas de aprendizaje automático, como se describe en la sección 4.6, son necesarias más de dos capturas de cada grafo, como mínimo se necesita un conjunto de entrenamiento, uno de validación y otro de test. Para el grafo de interacciones se resolvió haciendo dos particiones temporales más de los datos recogidos, una al final de la primera semana, y otra al final de la segunda. Obteniendo de esta forma 2 grafos más a partir de los cuales obtener los patrones y clases. Respecto al grafo de follows se utilizó como grafo de validación la captura en t_1 y como grafo de test una captura de las relaciones de seguimiento entre los usuarios del conjunto previo recogida el 20 de abril de 2017.

[4] Twitter REST API: https://dev.twitter.com/rest/public (Accedida 25/06/2017)

4.1.3 Configuración de las pruebas

Consideraremos la captura del estado de la subred de Twitter en el momento t_0 como el grafo de entrenamiento, y la captura del momento t_1 como el grafo de test.

El grafo de entrenamiento será el utilizado para calcular todas las características y métricas sobre el grafo, sus vértices y sus enlaces. Además, las recomendaciones se elaborarán en relación a este grafo, es decir, se construirán rankings que incluyan predicciones sobre la desaparición de los enlaces existentes entre los distintos usuarios del grafo de entrenamiento. Por su parte, el grafo de test será el utilizado para determinar si los citados enlaces se consideran como desaparecidos, si ya no están presentes en este grafo, o persistentes, si existen tanto en el grafo de entrenamiento como en el grafo de test. Los enlaces que aparecen en el grafo de test y que no aparecían previamente en el grafo de entrenamiento son despreciados.

La Figura 3 ejemplifica la clasificación previamente explicada de los enlaces, donde se observa que dicha clasificación se efectúa solo sobre el grafo de entrenamiento.

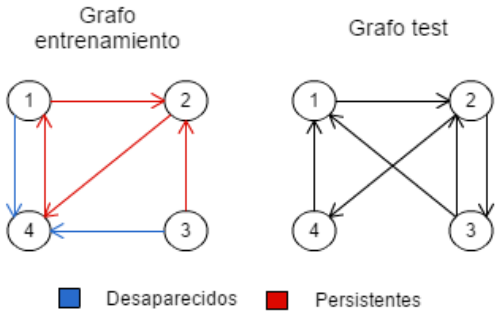

Figura 3. Clasificación de los enlaces en desaparecidos y persistentes.

Las pruebas se llevarán a cabo sobre los dos grafos mencionados en la sección 4.1.1 y 4.1.2, es decir, un grafo que representa las relaciones de seguimiento entre los distintos usuarios, una red social estable, al cual denominaremos *follows*, y otro grafo que representa una red social dinámica de interacciones, al cual nos referiremos por el nombre de *interacciones*.

4.2 Herramientas y código utilizado

Para la elaboración de este trabajo y la obtención de los resultados contenidos en el mismo se ha utilizado un repositorio proporcionado por el Grupo de Recuperación de Información de la Universidad Autónoma de Madrid, del que forma parte el tutor de este TFG. Este repositorio contiene código Java en su 8ª versión, que permite leer los ficheros de texto que contienen una lista con los enlaces entre pares de usuarios existentes en cada red social y transformarlos en una estructura de grafo. Dicho código contiene también las herramientas necesarias para, dado un sistema de puntuación de los usuarios que forman parte del extremo final de un enlace, generar un recomendador de ranking para todos los usuarios presentes. En cuanto a la evaluación de la calidad de las recomendaciones propuestas, el repositorio contiene diversos algoritmos que permiten calcularlas, pero por simplicidad y considerarse suficientemente ilustrativos solo se han utilizado los algoritmos de precisión y recall para la evaluación de recomendaciones en este trabajo.

El repositorio proporcionado tiene como base la librería externa RankSys[5], un marco público para la implementación y evaluación de algoritmos de recomendación. Esta librería contiene diversos algoritmos de recomendación y de evaluación que sirven de base para la extensión que el grupo de investigación ha realizado para adaptarlo a las características particulares del problema.

En cuanto a la parte de desarrollo e implementación propia se encuentra la codificación de todas las métricas de análisis de redes sociales descritas en la sección 4.3, así como la elaboración de los recomendadores de la sección 4.4, que utilizan dichas métricas para otorgar una puntuación a cada usuario objeto de estudio. De igual manera, la clasificación de los enlaces existentes en los grafos entre desaparecidos y persistentes y el estudio de las medias y distribuciones de los valores de las métricas ha supuesto la creación completa de código por cuenta personal.

Respecto a los métodos de predicción de aparición de enlaces contenidos en la sección 4.5, fueron proporcionados todos a excepción del de Vecinos Comunes, para que sirviesen además como modelo de cómo generar un recomendador propio. La creación de todos los métodos invertidos corrió por cuenta propia.

Como se ha comentado previamente los algoritmos de precisión y recall venían contenidos en el repositorio, aunque hubo que hacer diversas modificaciones en la función de relevancia que utilizaban para determinar si los usuarios recomendados a otro usuario eran relevantes o no, ya que dicha función estaba originalmente diseñada para el problema de predicción de enlaces y no para el de desaparición. Por tanto, en lugar de clasificar una recomendación como relevante si se encuentra presente en el grafo de test, el conjunto de recomendaciones relevantes se considera aquel en el que el enlace a evaluar se encuentra entre los enlaces desaparecidos del grafo, conjunto calculado previamente.

Respecto a la parte de aprendizaje automático abordada en la sección 4.6, la base de datos que contenía toda la información sobre los usuarios utilizada fue proporcionada por el grupo de investigación, al igual que las particiones temporales de los enlaces necesarias para crear los ficheros de entrenamiento y test de los modelos de aprendizaje automático. Sin embargo, la creación de los archivos de entrenamiento y test de entrada a modelos de aprendizaje automático y que contenían toda la información pertinente junto a la categorización de la clase de los enlaces en persistentes o desaparecidos fue trabajo mío.

Como herramientas utilizadas durante este proyecto destacan el entorno de programación Eclipse para toda la parte de codificación, elaborada en código Java y el uso de la plataforma de aprendizaje automático y minería de datos Weka. Para la elaboración de las gráficas y tablas que recogen todos los resultados obtenidos de los experimentos se ha utilizado la herramienta Excel de Microsoft Office.

[5] RankSys: http://ranksys.org (Accedida 25/06/2017)

4.3 Análisis de los enlaces de los grafos

Antes de comenzar a elaborar recomendadores basados en distintos algoritmos o métricas propias de las redes sociales, procedemos a hacer un análisis de cómo varían entre los enlaces desaparecidos y persistentes ciertas características o métricas propias del estudio de redes sociales.

4.3.1 Valores medios de métricas de redes sociales

En primer lugar, estudiamos los valores medios que toman dichas métricas sobre los distintos enlaces del grafo. En concreto las métricas que se han estudiado son, betweenness (sin normalizar), arraigo y reprocidad como métricas propias de los enlaces y por otro lado el grado incidente y saliente, betweenness (sin normalizar), coeficiente de clustering y el PageRank tanto de los vértices origen y destino de cada enlace, ya que también aportan información relevante sobre el enlace en cuestión. La explicación detalla de cada métrica y la fórmula que permite calcularla se encuentra en la sección 2.2.1.

En la Figura 4 y la Figura 5 podemos observar gráficas que contienen una comparación de todos los valores medios de las métricas mencionadas para los enlaces del grafo de follows y el grafo de interacciones respectivamente. Para cada métrica, la barra de menor tamaño representa el porcentaje que el valor de la métrica para ese tipo de enlaces supone respecto al valor mayor, que corresponde al otro tipo de enlaces.

Analizando ambas figuras desde una perspectiva general se aprecia que los valores de las métricas son mayoritariamente más bajos para los enlaces que desaparecen de la red.

Enlaces con menor arraigo, es decir, enlaces entre usuarios que comparten un menor número de amigos en común tienden a eliminarse. De igual manera lo hacen aquellos enlaces no recíprocos, lo cual tiene sentido, ya que es objetivamente más probable que un enlace desde el usuario v_1 al usuario v_2 desaparezca si el usuario v_2 no ha establecido a su vez un enlace hacia v_1, ya que la conexión entre ambos está menos reforzada.

Los enlaces que desaparecen también se caracterizan por estar establecidos entre personas con menor grado, tanto entrante (*indegree*) como saliente (*outdegree*), esto se debe simplemente a que la gran parte de las conexiones eliminadas son entre usuarios con un perfil estándar en cuanto a popularidad, desapareciendo pocas conexiones en las que están implicados usuarios con un gran número de seguidores (*followers*) o seguidos (*followees*).

Los usuarios que sirven de paso entre un menor número de nodos (valor de betweenness bajo) son también los más propensos a eliminar sus enlaces salientes o a que otro usuario elimine su conexión hacia ellos.

Los valores relativos al coeficiente de clustering de los nodos origen y destino de un enlace no presenta diferencias significativas entre los enlaces desaparecidos y persistentes, por tanto, no podemos sacar ninguna conclusión para detectar si un enlace desaparecerá valorando la cohesión del entorno de un nodo, es decir, la medida en la que los vecinos de dicho nodo están conectados entre sí.

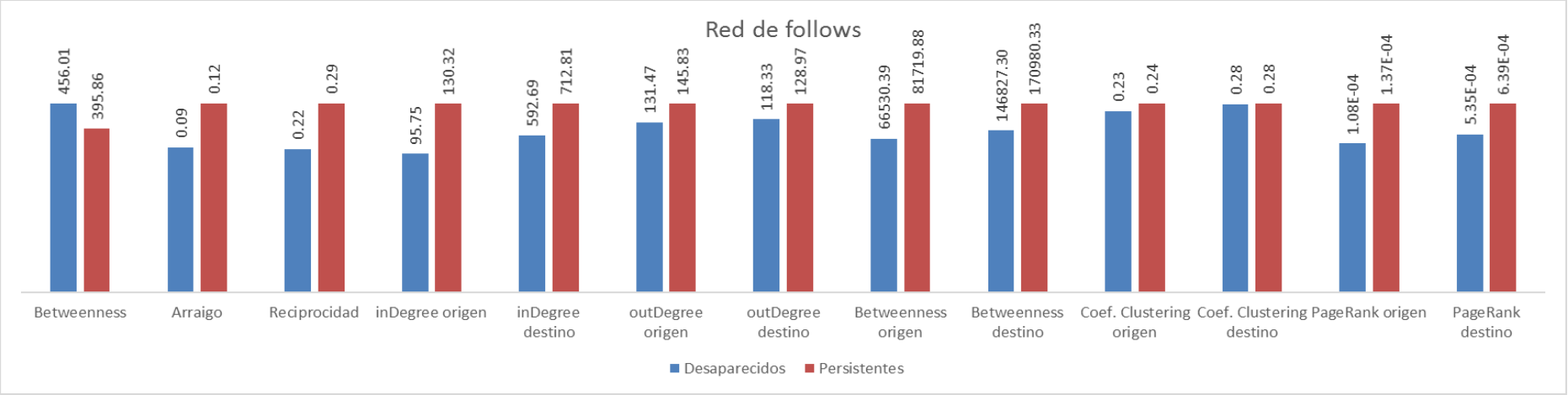

Figura 4. Valores medios de las métricas sobre los enlaces del grafo de follows, separados en desaparecidos y persistentes.

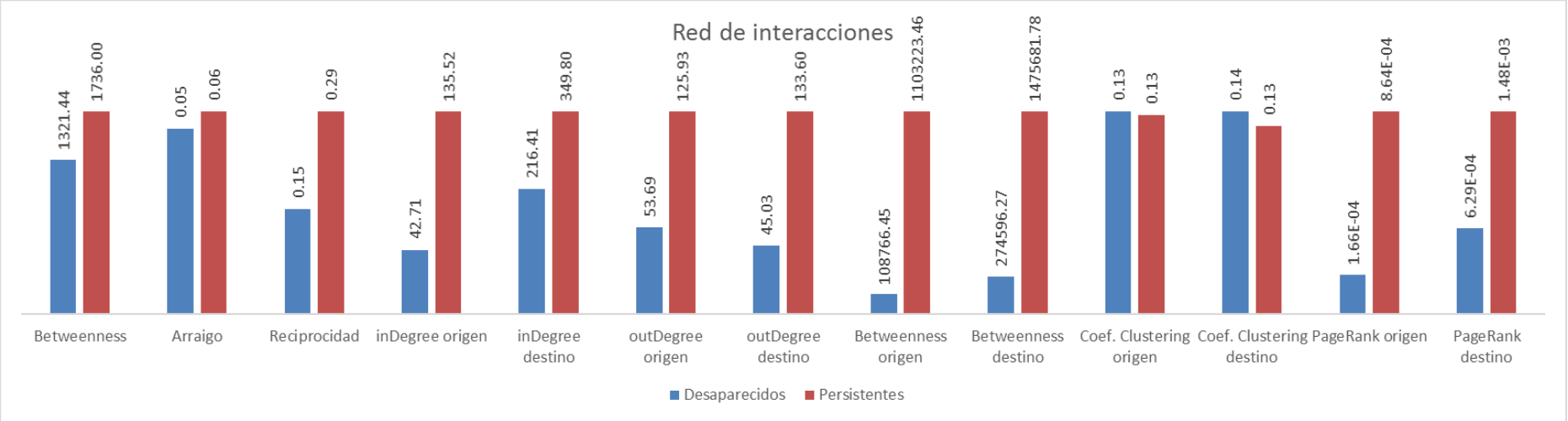

Figura 5. Valores medios de las métricas sobre los enlaces del grafo de interacciones, separados en desaparecidos y persistentes.

Respecto al PageRank de los distintos usuarios podemos sacar conclusiones similares, los usuarios menos relevantes según este algoritmo son los que están más implicados en el proceso de eliminación de conexiones, noción relacionada con lo anteriormente comentado para los usuarios de menor grado.

Sin embargo, esta simetría entre el grafo de Follows y de Interacciones no se cumple para el caso del betweenness de los enlaces. En el primer caso los enlaces desaparecidos fueron los que tenían un valor más alto de betweenness, es decir, aquellos enlaces "estratégicos" por los que pasa un gran número de caminos de distancia mínima. Sin embargo, en el caso del grafo de interacciones pasa justo lo contrario, los enlaces con menor influencia en el flujo de información, en otras palabras, aquellos con menor betweenness fueron los que más desaparecieron.

El último punto a destacar en este análisis es comparar la diferencia de proporciones que hay en el grafo de follows respecto al grafo de interacciones. En el grafo de follows la proporción de los valores de las métricas entre un tipo de enlaces y otros es aproximadamente del 80%, mientras que, en el grafo de interacciones, particularmente en los casos de la reprocidad, el grado y el PageRank de los nodos origen y destino del enlace, donde las diferencias son más significativas, ronda una media del 30%.

Todas las conclusiones obtenidas en esta sección son de gran importancia a la hora de elaborar un recomendador de enlaces con potencial de desaparición, ya que nos indican si es más conveniente ordenar los rankings elaborados de forma ascendente o descendente del valor de la métrica.

4.3.2 Distribución de las métricas

Si estudiamos la distribución de los valores de las métricas mencionadas en la sección 4.3.1, obtenemos una explicación adicional sobre cuál es la topología de los enlaces desaparecidos o persistentes de ambos grafos y nos ayudan a comprender mejor los valores medios obtenidos en la sección anterior.

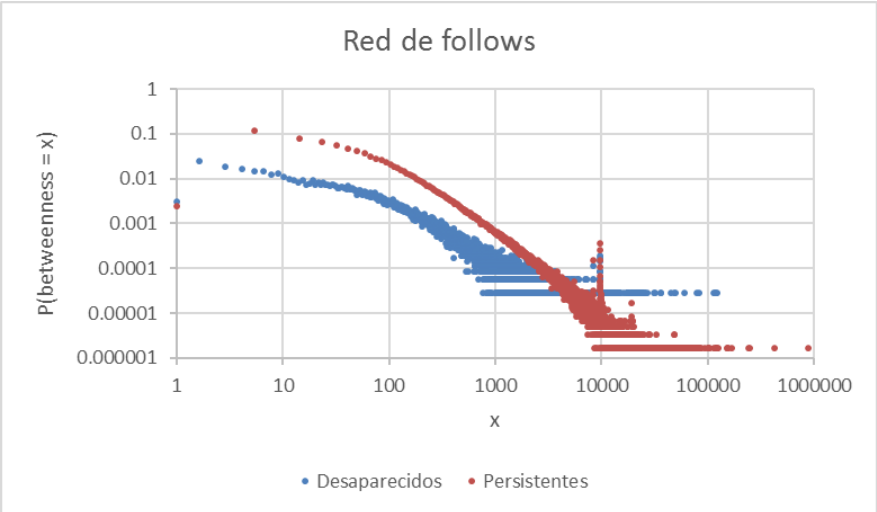

Figura 6. Distribución del betweenness de los enlaces del grafo de follows. Ejes en escala logarítmica.

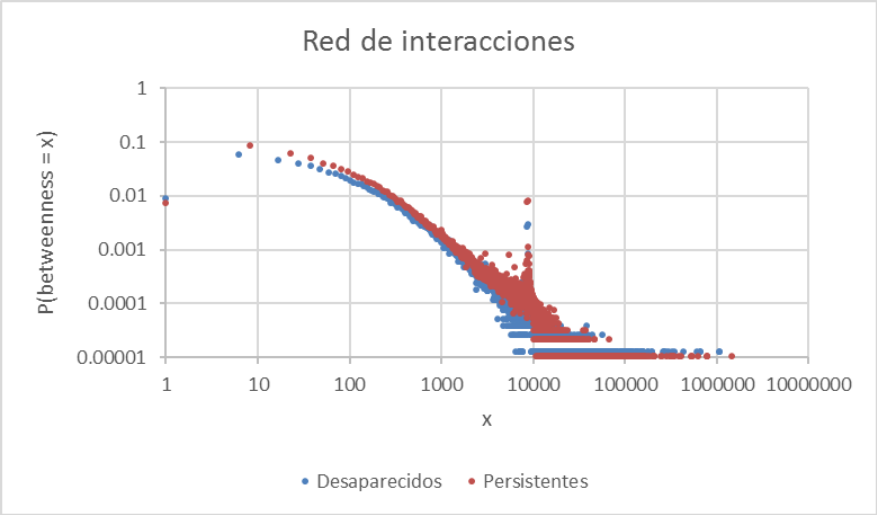

Figura 7. Distribución del betweenness de los enlaces del grafo de follows. Ejes en escala logarítmica.

La Figura 6 muestra la distribución del valor de betweenness para los enlaces persistentes y desaparecidos del grafo de follows, por otro lado, la Figura 7 ilustra la distribución de este valor en el grafo de interacciones. Comparando ambas gráficas se observa que ambas siguen una distribución power law, y que en el grafo de follows las diferencias entre ambos enlaces son superiores, mientras en el de interacciones, ambos enlaces tienen una distribución similar, con valores ligeramente más altos para el caso de los enlaces persistentes, lo que explica una media superior, como se vio en la sección 4.3.1.

Sin embargo, vemos que en el grafo de follows de la Figura 6 la distribución para cada tipo de enlaces difiere ligeramente, hay una gran cantidad de enlaces persistentes con valores bastante bajos (por debajo de 100), aunque luego existen una mínima cantidad de enlaces persistentes con valores elevados. Por su parte los enlaces desaparecidos se encuentran más repartidos entre valores intermedios, lo que acaba resultando en una media mayor del betweenness para estos últimos.

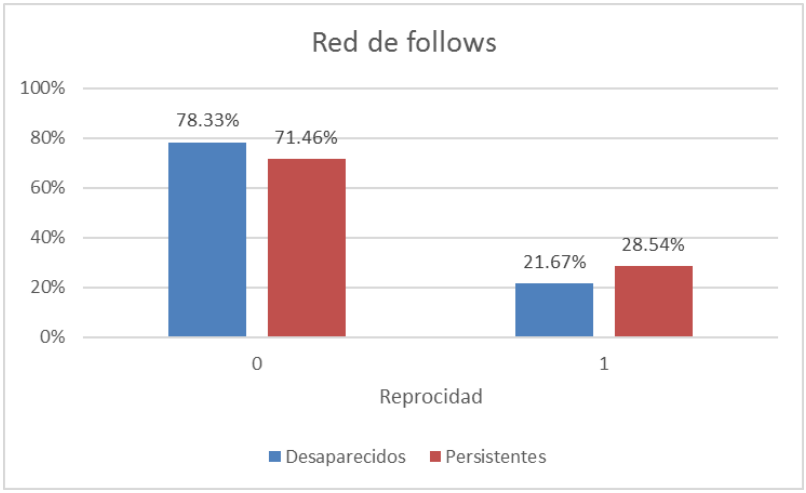

Figura 8. Distribución de la reprocidad de los enlaces del grafo de follows.

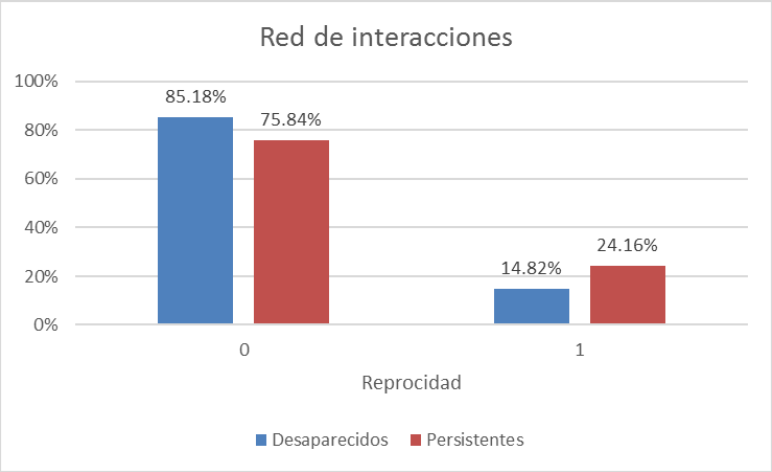

Figura 9. Distribución de la reprocidad de los enlaces del grafo de interacciones.

En relación a la reprocidad de los enlaces podemos observar que tanto las distribuciones de la Figura 8 y de la Figura 9 muestran un comportamiento similar. Predominan claramente en el grafo los enlaces no recíprocos (aquellos que tienen asociado un valor 0 de reciprocidad), para ambos tipos de enlaces, desaparecidos como persistentes. Aunque se puede apreciar un ligero mayor número de enlaces desaparecidos no recíprocos que de enlaces no recíprocos persistente, lo que da lugar a una media más baja de reciprocidad para estos enlaces.

Por último, observamos las gráficas de la Figura 10 y de la Figura 11 que representan las distribuciones del coeficiente de clustering del nodo destino del enlace, es decir, del nodo desde al que llega el enlace dirigido que sale desde el nodo que denominamos nodo origen. En ambas gráficas vemos que las distribuciones asociadas a los enlaces desaparecidos y persistentes son casi indistinguibles, lo que daría unos valores medios casi idénticos como los observados en la sección 214.3.1 y tienen la forma de una binomial. Comparando ambas figuras observamos que los valores con probabilidades más altas se encuentran acotados en el grafo de interacciones entre los coeficientes de clustering 0 y 0.2 mientras que en el grafo de follows la distribución es notablemente más ancha y baja y distribuye estos valores entre 0 y 0.4.

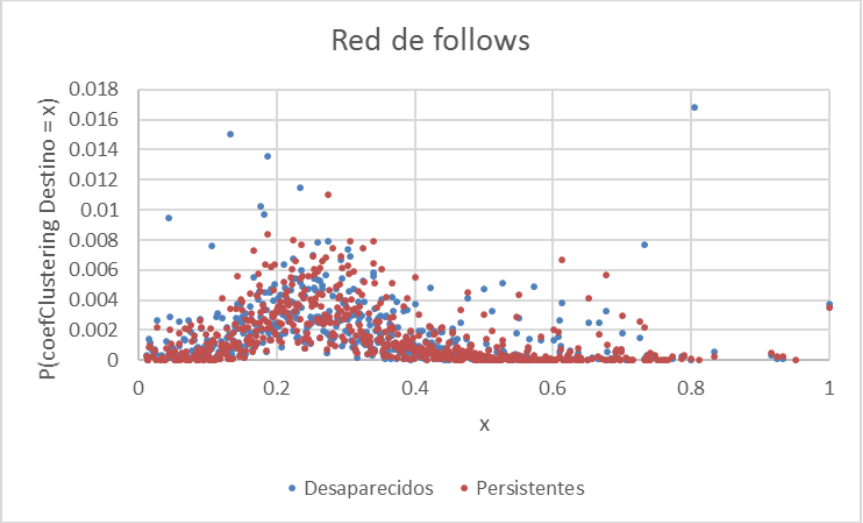

Figura 10. Distribución del coeficiente de clustering del nodo destino de los enlaces del grafo de follows.

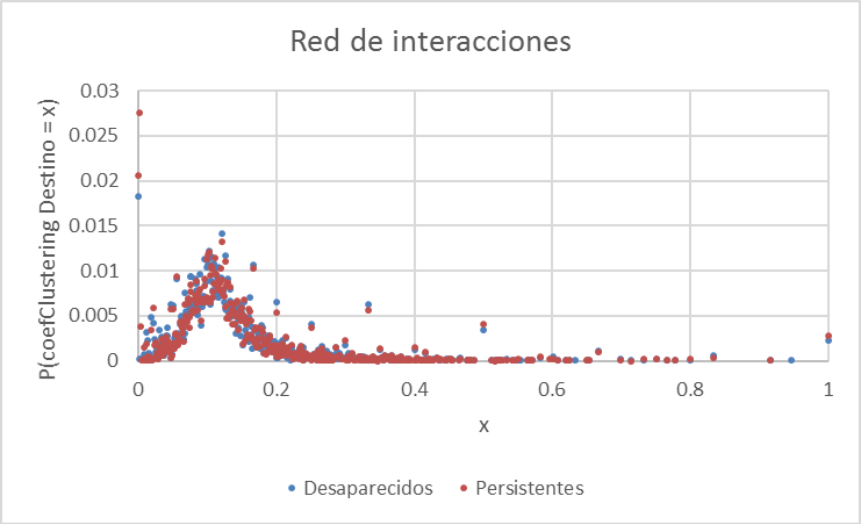

Figura 11. Distribución del coeficiente de clustering del nodo destino de los enlaces del grafo de interacciones.

La colección completa de gráficas con las distribuciones de las métricas estudiadas sobre los enlaces desaparecidos y persistentes del grafo de follows y el grafo de interacciones se puede consultar en el Anexo A. Distribuciones de métricas].

4.4 Métodos de recomendación por métricas de análisis de redes sociales

En relación a las métricas analizadas en la sección 4.3, el objetivo de este apartado es estudiar la calidad de recomendadores basados en ranking que utilizan métricas de análisis de redes sociales para asignar una puntuación a todos los usuarios con los que un usuario dado tiene una conexión establecida. Para ello, se calcula el valor de una determinada métrica para el enlace entre ambos usuarios o para el usuario que representa el nodo destino del enlace, en el caso de métricas de estudio de nodos.

La Tabla 1 recoge los valores de la evaluación de los recomendadores que utilizan las siguientes métricas: betweenness de los enlaces, betweenness del nodo destino de un enlace, arraigo y reprocidad de los enlaces, el grado de entrada y salida del nodo destino, así como el coeficiente de clustering y el PageRank del nodo destino.

Con el objetivo de determinar la calidad de las recomendaciones ofrecidas a los usuarios, se ha evaluado la precisión y recall de las mismas para distintos k, obteniendo de esta manera la utilidad para el usuario que se le ofrecen en el top k de la lista. Estos valores se comparan con la evaluación de un recomendador aleatorio, considerándose como buenos recomendadores aquellos que superan los valores de calidad de este recomendador.

Se observa que el comportamiento de los enlaces desaparecidos estudiado en la sección 4.3 determina la calidad de los distintos recomendadores. Por tanto, a excepción de los recomendadores por betweenness y reprocidad de la conexión entre los usuarios y el de coeficiente de clustering, el resto de recomendadores verifican que en su versión estándar obtienen resultados peores que el recomendador aleatorio, y en su versión invertida resultados mejores para cualquiera de las técnicas de evaluación consideradas.

Tabla 1. Evaluación de recomendadores basados en métricas de análisis de redes sociales sobre el grafo de follows.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.0664	0.0669	0.0657	0.0623	0.0184	0.0538	0.0843	0.1483
Betweenness enlace	0.0753	0.0718	0.0688	0.0660	0.0207	0.0550	0.0854	0.1555
Betweenness enlace invertido	0.0710	0.0655	0.0652	0.0634	0.0221	0.0537	0.0867	0.1559
Betweenness	0.0605	0.0588	0.0569	0.0548	0.0146	0.0406	0.0650	0.1181
Betweenness invertido	0.0869	0.0867	0.0853	0.0770	0.0266	0.0792	0.1262	0.2066
Arraigo	0.0544	0.0537	0.0527	0.0511	0.0156	0.0416	0.0650	0.1183
Arraigo invertido	0.0982	0.0890	0.0836	0.0761	0.0326	0.0784	0.1144	0.1879
Reciprocidad	0.0912	0.0763	0.0710	0.0653	0.0292	0.0685	0.1022	0.1756
Reciprocidad invertido	0.1198	0.0997	0.0923	0.0816	0.0404	0.0881	0.1278	0.2059
InDegree	0.0505	0.0541	0.0538	0.0523	0.0120	0.0368	0.0585	0.1060
InDegree invertido	0.0808	0.0793	0.0774	0.0722	0.0254	0.0697	0.1104	0.1918
OutDegree	0.0633	0.0621	0.0609	0.0589	0.0165	0.0467	0.0751	0.1360
OutDegree invertido	0.0841	0.0814	0.0818	0.0759	0.0245	0.0687	0.1142	0.1948
Coeficiente Clustering	0.1201	0.0890	0.0777	0.0689	0.0404	0.0775	0.1061	0.1720
Coeficiente Clustering invertido	0.0854	0.0787	0.0744	0.0677	0.0260	0.0669	0.1009	0.1706
PageRank	0.0510	0.0549	0.0546	0.0532	0.0119	0.0368	0.0595	0.1081
PageRank invertido	0.0822	0.0803	0.0777	0.0721	0.0268	0.0722	0.1117	0.1928

Consideramos como versión estándar de un recomendador aquella en la que el ranking de usuarios con los que un usuario dado va a perder potencialmente una conexión está ordenado en orden descendente de puntuación. Por el contrario, la versión invertida sería aquella en la que el ranking está ordenado en orden ascendente.

De esta forma, por ejemplo, para el recomendador por betweenness de los enlaces, el recomendador estándar sugeriría al usuario eliminar las conexiones con los usuarios cuyo enlace tuviese un betweenness más alto entre todos los enlaces que tiene establecidos un usuario hace otros usuarios de la red social en cuestión. Mientras que el recomendador invertido le recomendaría eliminar a aquellos usuarios cuya conexión tuviese valores más bajos de betweenness.

Se ha comentado previamente que el recomendador que utiliza el coeficiente de clustering de los usuarios finales de los enlaces evaluados, y los que utiliza el betweenness de las conexiones y su reprocidad no tienen el mismo comportamiento que el resto de recomendadores. Esto se debe a lo que ya se detectó en la sección 4.3, en el caso del coeficiente de clustering los valores medios de esta métrica para los dos tipos de enlaces eran esencialmente el mismo y sus distribuciones prácticamente indistinguibles, observamos sin embargo que como predictor de enlaces que van a desaparecer funciona mejor el recomendador estándar, logrando un mayor número de aciertos con los usuarios destino que tienen coeficientes de clustering mayores, a pesar de que el recomendador invertido también mejora al recomendador aleatorio. Por otro lado, respecto al recomendador por reprocidad de los enlaces se mantiene el comportamiento general de obtención de mayor calidad de

recomendaciones para la versión invertida, pero la versión estándar consigue unos resultados aceptables también. Por último, el recomendador que utiliza el valor de betweenness de los enlaces notamos que actúa de manera similar al del coeficiente de clustering, consigue superar al recomendador aleatorio en ambas de sus versiones, con resultados ligeramente mejores para la versión estándar.

Destacar que el recomendador de mayor calidad, considerando tanto precisión como recall, en la red social estable (grafo de follows) es el recomendador invertido que utiliza la reprocidad de los enlaces. Esto significa que los usuarios tienden a eliminar la conexión que han establecido con usuario que ha decidido no establecer a su vez una conexión con ellos, en otras palabras, es más probable que un usuario deje de seguir a otro usuario si este no le sigue.

Otros recomendadores con resultados destacados son los que utilizan el grado entrante o saliente de los usuarios considerados, y los que están basados en el betweenness y el PageRank de esos usuarios, todos ellos en su versión invertida. Esto nos indicaría que la tendencia de eliminación de conexiones, y en este caso la tendencia para dejar de seguir a un usuario, acción denominada *unfollow* suele producirse mayoritariamente hacia aquellos usuarios no demasiado populares, es decir, los que cuentan con un número bajo de *followers* (valores de indegree pequeños) o con un número no muy elevado de personas a las que siguen, followees, (valores de outdegree bajos). También aquellos usuarios menos clave en el paso de información son eliminados (valors de betweenness bajos), al igual que aquellos que no son considerados demasiado relevantes según el algoritmo PageRank.

Tabla 2. Evaluación de recomendadores basados en métricas de análisis de redes sociales sobre el grafo de interacciones.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.4314	0.3992	0.3720	0.3148	0.1290	0.3004	0.4049	0.5502
Betweenness enlace	0.3283	0.3759	0.3642	0.3157	0.0984	0.2909	0.4015	0.5505
Betweenness enlace invertido	0.4381	0.3973	0.3680	0.3116	0.1411	0.3064	0.4071	0.5500
Betweenness	0.2687	0.3424	0.3378	0.2979	0.0893	0.2798	0.3890	0.5394
Betweenness invertido	0.4580	0.4230	0.3901	0.3282	0.1453	0.3145	0.4173	0.5610
Arraigo	0.4253	0.3815	0.3545	0.3006	0.1349	0.2966	0.3989	0.5405
Arraigo invertido	0.4515	0.4182	0.3876	0.3243	0.1336	0.3042	0.4109	0.5555
Reciprocidad	0.3224	0.3525	0.3429	0.3034	0.0987	0.2823	0.3919	0.5437
Reciprocidad invertido	0.4633	0.4256	0.3917	0.3280	0.1466	0.3178	0.4204	0.5613
InDegree	0.2863	0.3329	0.3259	0.2913	0.0924	0.2747	0.3823	0.5342
InDegree invertido	0.4856	0.4387	0.4028	0.3360	0.1504	0.3221	0.4248	0.5661
OutDegree	0.2683	0.3456	0.3387	0.2990	0.0899	0.2819	0.3892	0.5391
OutDegree invertido	0.4553	0.4193	0.3864	0.3270	0.1459	0.3144	0.4160	0.5608
Coeficiente Clustering	0.4704	0.4261	0.3886	0.3232	0.1428	0.3148	0.4152	0.5551
Coeficiente Clustering invertido	0.4057	0.3878	0.3628	0.3084	0.1174	0.2896	0.3973	0.5450
PageRank	0.2734	0.3384	0.3337	0.2946	0.0896	0.2764	0.3858	0.5361
PageRank invertido	0.4833	0.4391	0.4021	0.3345	0.1501	0.3224	0.4245	0.5644

Unos resultados similares se obtienen cuando analizamos la calidad de las recomendaciones de estos mismos recomendadores en el grafo de interacciones, los cuales se recogen en la Tabla 2. En este caso todos los recomendadores obtienen mejores resultados en sus versiones invertidas, a excepción del que utiliza el coeficiente de clustering de los usuarios objetivo, lo que es coherente con las medias observadas para estas métricas en la Figura 5.

Se observa que recomendador invertido por betweenness de los enlaces no es demasiado fiable, ya que su calidad es bastante similar a la del recomendador aleatorio y que por otra parte el recomendador que utiliza el coeficiente de clustering, al igual que en el grafo de follows, vuelve a ser más eficaz en su versión estándar.

Y de forma similar a lo que ocurría en el grafo de follows y que se observaba en la Tabla 1, en el grafo de interacciones destacan los recomendadores basados en los grados entrantes y salientes y el basado en el PageRank, seguidos por el de reprocidad, todos ellos en su versión invertida. Lo que ratifica su papel como mejores recomendadores de enlaces que van a desaparecer en un grafo asociado a una red social.

Para concluir esta sección comentar el porqué de esa gran diferencia de precisión que se consigue en la red de follows frente a la red de interacciones. Esto se debe a la proporción de enlaces que desaparecen en ambos grafos. Como vimos en la sección 4.1.1, solo el 5.65% de los enlaces desaparecen en el grafo de follows, mientras que en el de interacciones este valor se eleva al 45.26%, valores que se aproximan a los resultados que consigue el recomendador aleatorio. Por tanto, será mucho más probable predecir correctamente que un enlace va a desaparecer en la red de interacciones debido a que existe una mayor cantidad de enlaces con esta propiedad de cambio de estado.

4.5 Inversión métodos de predicción de aparición de enlaces

Existen numerosos métodos de predicción de enlaces en redes sociales, en este trabajo nos centraremos en el desarrollo de un grupo reducido de los mismos, tal como se indicó en la sección 3.2, por considerarse suficientemente representativos de este gran abanico de métodos.

Para ello se ha analizado el resultado obtenido por los recomendadores por ranking en su versión estándar e invertida, según se comentó en la sección 4.4, utilizando en este caso métodos de predicción de aparición de enlaces para redes sociales.

Como se puede apreciar en la Tabla 3 y en la Tabla 4, los métodos de popularidad, Adamic-Adar, recomendación basada en vecindarios, amigos comunes y BM25 no obtienen resultados satisfactorios en su versión estándar, ya que originalmente se propusieron como predictores de aparición de enlaces, y por tanto como cabría esperar no son buenos detectores de enlaces con potencial de desaparición.

Sin embargo, podemos observar en ambas tablas que la inversión del sistema de puntuación de estos algoritmos ofrece recomendadores de desaparición de enlaces relativamente buenos, destacando por su calidad el recomendador invertido de vecinos comunes. La idea detrás de la inversión de este algoritmo reside en ofrecer una puntuación más baja para aquellos usuarios con los que no se comparta un gran conjunto de amigos, de forma que los usuarios con los que una persona dada tenga una conexión, pero no tengan amigos en común o solo

un grupo reducido de ellos, obtengan una puntuación en la recomendación de desaparición de la conexión entre ambos más alta.

Tabla 3. Evaluación de recomendadores de predicción de aparición de enlaces y sus inversos sobre el grafo de follows.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.0664	0.0669	0.0657	0.0623	0.0184	0.0538	0.0843	0.1483
Popularidad	0.0505	0.0541	0.0538	0.0523	0.0120	0.0368	0.0585	0.1060
Popularidad invertido	0.0808	0.0793	0.0774	0.0722	0.0254	0.0697	0.1104	0.1918
Adamic-Adar	0.0424	0.0455	0.0452	0.0460	0.0095	0.0298	0.0477	0.0915
Adamic invertido	0.1174	0.0995	0.0937	0.0834	0.0434	0.0971	0.1423	0.2289
User-based kNN	0.0439	0.0449	0.0448	0.0450	0.0101	0.0297	0.0478	0.0911
User-based kNN invertido	0.0906	0.0871	0.0838	0.0761	0.0285	0.0788	0.1192	0.1982
Vecinos comunes	0.0431	0.0454	0.0458	0.0464	0.0097	0.0299	0.0484	0.0922
Vecinos comunes invertido	0.1196	0.1006	0.0950	0.0837	0.0449	0.0998	0.1473	0.2313
BM25	0.0608	0.0585	0.0566	0.0550	0.0149	0.0405	0.0645	0.1196
BM25 invertido	0.1140	0.0962	0.0899	0.0786	0.0421	0.0912	0.1307	0.2047

En otras palabras, a raíz de los resultados obtenidos podemos determinar que los patrones que favorecen que un usuario elimine de su red de conexiones a otro usuario, tanto en relación a dejar de seguir a dicho usuario como a interrumpir las interacciones con el mismo, está ligado a que dicho usuario no sea muy popular (bajo número de seguidores, lo que se corresponde con el recomendador por grado incidente invertido estudiado en la sección 4.4). También favorece la desaparición de un enlace que los usuarios con los que se encuentran conectados ambos usuarios del enlace evaluado cuenten con un gran número de vecinos (en el caso de Adamic-Adar), a la compartición de pocas características en común (User-based kNN), a un número reducido de amigos en común y a la baja relevancia del enlace según el algoritmo BM25.

Tabla 4. Evaluación de recomendadores de predicción de aparición de enlaces y sus inversos sobre el grafo de interacciones.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.4314	0.3992	0.3720	0.3148	0.1290	0.3004	0.4049	0.5502
Popularidad	0.2863	0.3329	0.3259	0.2913	0.0924	0.2747	0.3823	0.5342
Popularidad invertido	0.4856	0.4387	0.4028	0.3360	0.1504	0.3221	0.4248	0.5661
Adamic-Adar	0.2737	0.3201	0.3156	0.2814	0.0996	0.2713	0.3784	0.5290
Adamic invertido	0.4950	0.4502	0.4128	0.3431	0.1469	0.3249	0.4297	0.5713
User-based kNN	0.2446	0.3035	0.2977	0.2677	0.0852	0.2553	0.3495	0.4865
User-based kNN invertido	0.4938	0.4385	0.3979	0.3284	0.1463	0.3070	0.3995	0.5270
Vecinos comunes	0.2982	0.3260	0.3179	0.2826	0.1086	0.2742	0.3798	0.5298
Vecinos comunes invertido	0.4935	0.4489	0.4116	0.3420	0.1455	0.3236	0.4284	0.5709
BM25	0.4093	0.3813	0.3561	0.3030	0.1290	0.2930	0.3971	0.5422
BM25 invertido	0.4787	0.4273	0.3888	0.3262	0.1395	0.3103	0.4126	0.5567

Comparando la calidad de las recomendaciones en la red social estable y en la de interacciones no observamos diferencias significativas de calidad, aunque la proporción de mejora respecto al recomendador aleatorio es mayor en el grafo de follows, es decir, en la red social estable.

4.6 Adaptación predicciones aprendizaje automático supervisado al problema de recomendación

Otra de las aproximaciones al problema de recomendación que se ha explorado es la posible adaptación de predicciones obtenidas a partir de métodos de aprendizaje automático supervisado basados en características propias de los usuarios.

Para ello lo primero que se ha hecho es construir los ficheros de entrenamiento y de test, combinando en los ficheros distintas particiones temporales del grafo de follows y de interacciones. Como se explicó en la sección 4.1.1 para el caso del grafo de follows contamos con tres ficheros con distintas capturas temporales de las relaciones de seguimiento entre un conjunto de usuarios concreto, lo que haremos entonces es generar el fichero de entrenamiento combinando las dos primeras capturas, el fichero 1 sirve para establecer los datos o patrones de las instancias (enlaces) y de la comparación entre el fichero 1 y el 2 obtenemos las clases para este primer fichero de entrenamiento. Realizamos el mismo proceso entre los ficheros 2 y 3, en este caso el fichero 2 (fichero de validación) sirve para establecer los patrones y las clases que se registran en el fichero de test. En el caso del grafo de interacciones, al contar con 4 capturas temporales del mismo la generación de los ficheros de entrenamiento y test se hace comparando el fichero 1 y 2 para generar el fichero de entrenamiento y por otra parte los ficheros 3 y 4 para generar el fichero de test. La Figura 12 y la Figura 13 ilustran mejor este proceso.

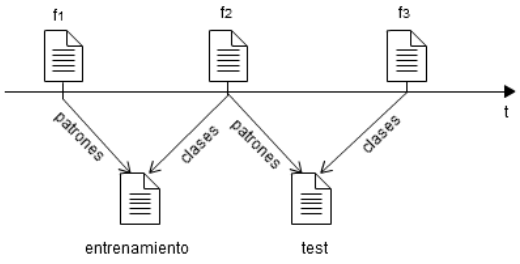

Figura 12. Proceso de generación ficheros de entrenamiento y test en el grafo de follows.

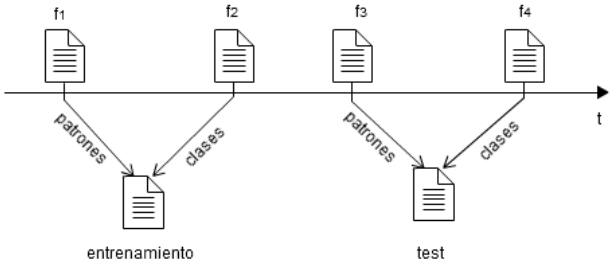

Figura 13. Proceso de generación ficheros de entrenamiento y test en el grafo de interacciones.

Las características que se han considerado de los usuarios de Twitter involucrados en los enlaces analizados son si disponen de una cuenta verificada o no, su número de seguidores, el número de personas a las que sigue, número de listas en las que aparece el usuario, número de tweets que ha publicado desde que se creó la cuenta y número de tweets que ha publicado mientras se hacía la recogida de información para obtener el conjunto de datos.

El fichero de prueba y de test generado para los dos tipos de grafos distintos objeto de nuestro estudio son sometidos, gracias a la ayuda del programa Weka, a tres algoritmos distintos de aprendizaje automático, Naive Bayes, Regresión logística y Bosques aleatorios con los parámetros por defecto. Los cuales permitirán, basándose en los datos de las instancias

contenidas en el fichero de entrenamiento y la clase asociada a las mismas (Persistente o Desaparecido), clasificar los enlaces contenidos en test en una de estas dos clases.

Entre la información que obtenemos como salida al ejecutar los distintos clasificadores, nos encontramos con la denominada matriz de confusión, que tiene una estructura como la que se puede observar en la Figura 14, y que nos permite contabilizar el número de verdaderos positivos y negativos, así como los falsos positivos y negativos.

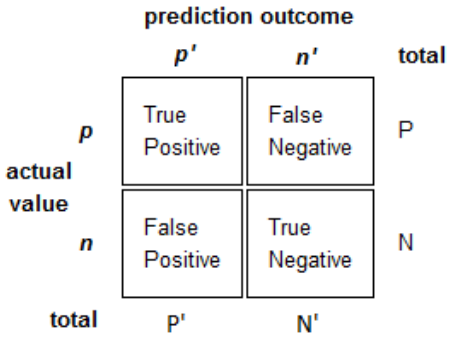

Figura 14. Matriz de confusión

Las matrices de confusión obtenidas para el grafo de follows son las que se ilustran en la Figura 15, la Figura 16 y la Figura 17. Podemos observar que Naive Bayes no consigue clasificar correctamente ninguno de los enlaces desaparecidos en el grafo, el bosque aleatorio una mínima parte de ellos, un 0.37%, pero sin embargo la regresión logística acierta con la clasificación del 38.94% de los enlaces desaparecidos.

```
     a      b   <-- classified as
562820      2 |      a = Persistente
 82200      0 |      b = Desaparecido
```

Figura 15. Matriz de confusión clasificador Naive Bayes en el grafo de follows.

```
     a      b   <-- classified as
372545 190277 |      a = Persistente
 50188  32012 |      b = Desaparecido
```

Figura 16. Matriz de confusión clasificador regresión logística en el grafo de follows.

```
     a      b   <-- classified as
562458    364 |      a = Persistente
 81895    305 |      b = Desaparecido
```

Figura 17. Matriz de confusión clasificador bosque aleatorio en el grafo de follows.

Estos malos resultados se deben en gran parte a lo que ya se comentó anteriormente sobre la gran desproporción de enlaces desaparecidos frente a persistentes que hay en el grafo de follows, lo que hace que los clasificadores tiendan a asignar la clase “Persistente” a los enlaces, ya que tienen más probabilidades de acertar que asignando la contraria. En el caso de los grafos utilizados para aprendizaje automático observamos que hay 562822 enlaces persistentes y 82200 enlaces desaparecidos, es decir, los enlaces desaparecidos suponen solo un 12.74% del total.

Para convertir la salida de los clasificadores a los recomendadores por ranking objeto de estudio de este TFG, lo que hacemos es utilizar la distribución de probabilidades que asigna

cada recomendador a cada enlace de pertenecer a la clase b (enlaces desaparecidos) y utilizar dicha puntuación para ordenar en orden descendente los enlaces de cada usuario.

Los resultados de la evaluación de los recomendadores así construidos para cada algoritmo de aprendizaje automático son bastante buenos, como se observa en la Tabla 5. Vemos además que los resultados observados en las matrices de confusión no tienen un reflejo directo en la calidad de los recomendadores, ya que Naive Bayes, a pesar de no ser capaz de clasificar ningún enlace como desaparecido, obtiene resultados buenos en su versión recomendador. Por otro lado, el bosque aleatorio es la mejor opción con diferencia, a pesar de la baja proporción de enlaces desaparecidos que conseguía identificar.

Esto nos indica que, aunque no se consiga hacer una clasificación de calidad, la distribución de probabilidad que se le asigna a los enlaces de pertenecer a la clase de desaparecidos es relativamente buena cuando se compara con esos mismos valores y no frente a la distribución de probabilidad de pertenecer a la clase de persistentes.

Tabla 5. Evaluación de recomendadores generados a partir de algoritmos de aprendizaje automático sobre el grafo de follows.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.0664	0.0669	0.0657	0.0623	0.0184	0.0538	0.0843	0.1483
NaiveBayes	0.1797	0.1709	0.1660	0.1533	0.0353	0.0979	0.1495	0.2483
Logistic	0.1945	0.1793	0.1710	0.1571	0.0422	0.1066	0.1595	0.2630
RandomForest	0.2257	0.2069	0.1927	0.1712	0.0543	0.1286	0.1846	0.2904

Las mismas técnicas de aprendizaje automático son entrenadas con el grafo de interacciones, obteniéndose los resultados de clasificación que se observan en la Figura 18, la Figura 19 y la Figura 20. En este caso Naive Bayes clasifica correctamente el 94.02% de los enlaces desaparecidos, la regresión logística el 99.92% y el bosque aleatorio el 69.45% de los mismos.

En este caso observamos que hay 77141 enlaces desaparecidos y 93283 enlaces persistentes, y que por tanto el número de enlaces desaparecidos constituyen un 45.26% del total, por lo que el conjunto de datos se encuentra más balanceado.

```
    a     b   <-- classified as
72529  4612 |     a = Desaparecido
84654  8629 |     b = Persistente
```

Figura 18. Matriz de confusión clasificador Naive Bayes en el grafo de interacciones.

```
    a     b   <-- classified as
77082    59 |     a = Desaparecido
93179   104 |     b = Persistente
```

Figura 19. Matriz de confusión clasificador regresión logística en el grafo de interacciones.

```
    a     b   <-- classified as
53572 23569 |     a = Desaparecido
53905 39378 |     b = Persistente
```

Figura 20. Matriz de confusión clasificador bosque aleatorio en el grafo de interacciones.

Si convertimos ahora estas predicciones al formato recomendador por ranking enfocado al usuario, según se ha descrito para el grafo de follows, se obtienen los valores de evaluación recogidos en la Tabla 6, donde se observa que los bosques aleatorios consiguen de nuevo una recomendación excepcional, aunque en el caso del grafo de interacciones los recomendadores de Naive Bayes y la regresión logística consiguen solo valores de relevancia para los usuarios un poco superiores al recomendador aleatorio, en comparación con el grafo de follows que en el que conseguían mejoras más significativas.

Tabla 6. Evaluación de recomendadores generados a partir de algoritmos de aprendizaje automático sobre el grafo de interacciones.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.4314	0.3992	0.3720	0.3148	0.1290	0.3004	0.4049	0.5502
NaiveBayes	0.4673	0.4260	0.3915	0.3298	0.1440	0.3167	0.4189	0.5620
Logistic	0.4736	0.4315	0.3973	0.3313	0.1453	0.3200	0.4228	0.5643
RandomForest	0.7749	0.6549	0.5724	0.4373	0.2272	0.4195	0.5157	0.6342

Como conclusión a esta sección podemos afirmar que la transformación de algoritmos de aprendizaje automático a recomendadores de ranking utilizando información propia de los usuarios en Twitter es una bastante buena opción ya que se superan con creces los umbrales mínimos de calidad de recomendación, siendo la mejora más significativa sobre el grafo de follows.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

En conclusión, tras explorar distintas aproximaciones al problema de la predicción de la desaparición de enlaces en redes sociales sobre dos grafos de naturaleza distinta en Twitter, una red de follows y una red de interacciones, podemos concluir que en líneas generales los recomendadores que funcionan bien en un grafo obtienen también resultados satisfactorios en el otro grafo.

De forma general, como se puede observar en el Anexo B. Evaluación de los recomendadores utilizados], los recomendadores que han obtenido mejores resultados en ambos grafos, aunque en especial en el grafo de follows, son los obtenidos a partir de la adaptación de predicciones de aprendizaje automático supervisado basadas en características propias de los usuarios, destacando el clasificador que utiliza bosques aleatorios.

Respecto al resto de recomendadores, sobre la red estable de follows han obtenido una mejor evaluación en términos de precisión y recall a distintos k, el recomendador de vecinos comunes, seguido de los recomendadores basados en reciprocidad de los enlaces, Adamic-Alar y BM25, todos ellos en su versión invertida, ya que en su versión estándar están diseñados para predecir la aparición de enlaces. Los peores resultados se han obtenido en este grafo para los algoritmos en formato invertido de betweenness de los enlaces, que consigue solo resultado ligeramente superiores al recomendador aleatorio.

Respecto a la red dinámica de interacciones los recomendadores en versión invertida que han demostrado una mejor actuación en este grafo son Adamic-Adar, seguido de cerca por el algoritmo de vecinos comunes y el de popularidad o grado incidente. Por otra parte, respecto a los algoritmos en versión invertida destacar de nuevo como un mal recomendador al basado en el betweenness de los enlaces.

Se ha observado por otro lado que cuando se trata de la métrica de coeficiente de clustering de un nodo, el recomendador asociado a la misma funciona correctamente para el problema dado de recomendación en su versión estándar, ya que como se comentó en la sección 2.2.1 enlaces con un alto clustering suelen venir asociados a un bajo betweenness, luego si el recomendador de betweenness funciona mejor en su versión estándar el que utiliza el coeficiente de clustering lo hará en su versión invertida.

Como conclusión podemos decir que, de las tres aproximaciones dadas al problema, recomendadores utilizando métricas de análisis de redes sociales, recomendadores elaborados a partir de la inversión de algoritmos clásicos de recomendación de contactos y modelos de aprendizaje automático basados en características de los usuarios, las dos primeras han conseguido resultados positivos, pero destacan las recomendaciones conseguidas a partir de métodos de aprendizaje automático.

5.2 Trabajo futuro

El presente trabajo presenta numerosas posibilidades de ampliación y trabajo futuro. En esta sección se mencionan algunas de ellas.

Como ya se mencionó previamente, el campo de la recomendación de contactos en redes sociales es aún un campo abierto de investigación y una gran cantidad de algoritmos y técnicas de recomendación siguen en continuo desarrollo, por tanto, no cabe duda de que la exploración de nuevos métodos, más efectivos y de mayor calidad supone una gran vía de desarrollo y expansión en este ámbito.

En particular, la predicción de desaparición de enlaces es un área que no ha sido suficientemente estudiada aún y se encuentra en una fase muy joven de desarrollo e investigación. Por este motivo, la búsqueda de nuevas aproximaciones al problema, tanto viéndolo como la otra parte del problema clásico de recomendación de contactos, como avanzando por vías innovadores y originales ofrece aún muchas posibilidades.

Una ampliación inmediata del estudio realizado en este trabajo sería aumentar el número y variedad de métricas utilizadas para estudiar la naturaleza de los enlaces desaparecidos y persistentes, así como una ampliación de los métodos clásicos de recomendación y aprendizaje automático empleados, detectando cuales encuentran mejores resultados para el problema de la predicción de enlaces que van a desaparecer o cuales son más ineficaces. Se podría complementar además este trabajo combinando en un único recomendador varios de los métodos que resulten más eficaces con el objetivo de beneficiarse de los aspectos positivos de todos ellos, así como utilizando las métricas de análisis de redes sociales como características adicionales de los datos utilizados para el aprendizaje automático.

También podría ser de interés seguir estudiando el conjunto de usuarios elegido para este experimento, y comprobar si a largo plazo las predicciones sobre enlaces con potencial de desaparición se cumplen, ya que las ventanas temporales con los que se ha trabajado pueden no haber sido lo suficientemente amplias como para que los usuarios hayan llevado a cabo la eliminación de dichos contactos.

Por último, este estudio se ha centrado solo en la red social de Twitter, por lo que otra posibilidad recae en estudiar otras redes sociales como Facebook, Instagram o LinkedIn que poseen propiedades diferentes a Twitter y que por tanto los resultados observados en este estudio no tienen por qué ser válidos en dichas redes sociales.

Referencias

Al Hasan, M., Chaoji, V, Salem, S. and Zaki, M. Link prediction using Supervised Learning. *Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

Adamic, L.A., Adar, E. *Friends and Neighbors on the Web*. Social Networks, 25(3), July 2003, pp. 211-230.

Backstrom, L. and Kleinberg, J. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. *Proc. 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, Baltimore, U.S.A, February 2014, pp. 831-841.

Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd Edition. Addison Wesley, 2010.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual Web search engine. *Proc. 7th Annual International Conference on World Wide Web (WWW 1998)*, Brisbane, Australia, April 1998, pp. 107-117.

Castells, P., Hurley, N.J. and Vargas, S. Novelty and Diversity in Recommender Systems. *Recommender Systems Handbook.* In Ricci et al. (2015), pp. 881-918.

Cremonesi, P., Koren, Y., Turrin, R. Performance of Recommender Algorithms on Top-N Recommendation Tasks. Proceedinf of the 4th Annual International ACM Conference on Recommender Systems (RecSys 2010), Barcelona, Spain, September 2010, pp. 39-46.

Durkheim, E. *De la division du travail social: étude sur l'organisaion des sociétés supérieures*. Alcan, 1893.

Garimella, K., Weber, I. and Dal Cin, Sonia. From "I Love you babe" to "leave alone" – Romantic Relationship Breakups on Twitter. *Proc. 6th International Conference on Social Informatics (SocInfo 2014)*, Barcelona, Spain, November 2014, pp. 199-215.

Goodman, L.A. Snowball sampling. *Annals of Mathematical Statistics*, 31(1), 1961, pp. 148-170.

Guimerà, R. and Sales-Pardo, M. Missing and Spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106 (52), December 2009, pp. 22073-22078.

Guns, R. Generalizing link prediction: Collaboration at the University of Antwerp as a case study. *Proceedings of the American Society for Information Science and Technology*, 46 (1), 2009, pp. 1–15.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang. D., Zadeh, R. WTF: The Who to Follow Service at Twitter. Proceedings of the *22nd Annual International Conference on World Wide Web (WWW 2013)*, Rio de Janeiro, Brazil, May 2013, pp. 505-514.

Kivran-Swaine, F., Govindan, P. and Naaman, M. The Impact of Network Structure on Breaking Ties in Online Social Networks: Unfollowing on Twitter. *Proceedings of the 29th international conference on Human Factors in Computing Systems (CHI '11)*, Vancouver, Canada, May 2011, pp. 1101-1104.

Kwak, H., Chun, H. and Moon, S. Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter. *Proceedings of the 29th international conference on Human Factors in Computing Systems (CHI '11)*, Vancouver, Canada, May 2011, pp. 1091-1100.

Leskovec, J., Huttenlocher, D. and Kleinberg, J. Predicting Positive and Negative Links in Online Social Networks. *Proc. 19th Annual International Conference on World Wide Web (WWW 2010)*. Raleigh, U.S.A, April 2010, pp. 641-650.

Lichtenwalter, R.N., Lussier, J.T and Chawla, N.V. New Perspectives and Methods in Link Prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*, Washington, U.S.A, 2010, pp. 243-252.

Macek, B.E., Atzmüller, M., and Stumme, G. Predicting the stability of user interaction ties in Twitter. *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business (I-Know* 2014), Graz, Austria, September 2014, article no. 9.

Myers, S.A., Sharma, A., Gupta, P., Lin, J. Information Network or Social Network?: The Structure of the Twitter Follow Graph. *Proceedings of the 23rd Annual International Conference on World Wide Web (WWW 2014)*, Seoul, Korea, April 2014, pp. 493-498.

Newman, M.E.J. (2001A). Clustering and Preferential Attachment in Growing Networks. *Physical Review Letters E*, 64(025102), April 2001.

Newman, M. E. The structure and function of complex networks. *SIAM Review*, 2003, 45(2), pp. 167-256.

Pan, L., Zhou, T. and Lü, L. Predicting missing links and identifying spurious links via likelihood analysis. *Scientific Reports*, 6, March 2016, article no. 22955.

Quercia, D., Bodaghi, M. and Crowcroft, J. Loosing "Friends" on Facebook. *Proceedings of the 4th Annual ACM Web Science Conference (WebSci 2012)*, Evanston, U.S.A., June 2012, pp. 251-254.

Shani, G., Gunawardana, A., Evaluating Recommender Systems. *Recommender Systems Handbook*. In Ricci et al. (2015), pp. 265-308.

Sparck Jones, K., Walker, S., Roberton S.E. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. *Information Processing and Management 36*. February 2000, pp. 779-808 (part 1), pp. 809-840 (part 2).

Tönnies, F. Gemeinschaft und Gesellschaft. *Fues' Verlag*. 1887.

Verbeke, W., Martens, D. and Baesens, B. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14, 2014, pp. 431-446.

Wang, P., Xu, B., Wu, Y. and Zhou, X. Link Prediction in Social Networks: the State-of-the-Art. *Science China Information Sciences*, 58(1), January 2015, pp. 1-38.

Xu, B., Huang, Y., Kwak, H. and Contractor, N.S. Structures of Broken Ties: Exploring Unfollow Behavior on Twitter. *Proceedings of the 2nd ACM Conference on Computer Supported Cooperative Work (CSCW)*, San Antonio, U.S.A., February 2013, pp. 871-876.

Zeng, A. and Cimini, G. Removing spurious interactions in complex networks. *Physical Review E*, 85(3), March 2012, pp. 36101-36107.

Glosario

Enlace	Conexión que se establece entre dos vértices de un grafo.
Grafo dirigido	Grafo cuyos enlaces comienzan en un nodo origen e inciden en un nodo destino.
Métrica	Fórmula o método que permite calcular una propiedad de un elemento.
Nodo	Unidad elemental de un grafo.
Ranking	Lista ordenada de resultados siguiendo un orden ascendente o descendente del valor de los mismos.
Recomendación	Sugerencia que se le hace a un usuario sobre la realización de una acción.
Relevancia	Utilidad que tiene para el usuario un elemento que se le recomienda.
Red Social	Red de personas que establecen interacciones entre ellas.

Anexo A. Distribuciones de métricas

A.1 Betweenness enlaces

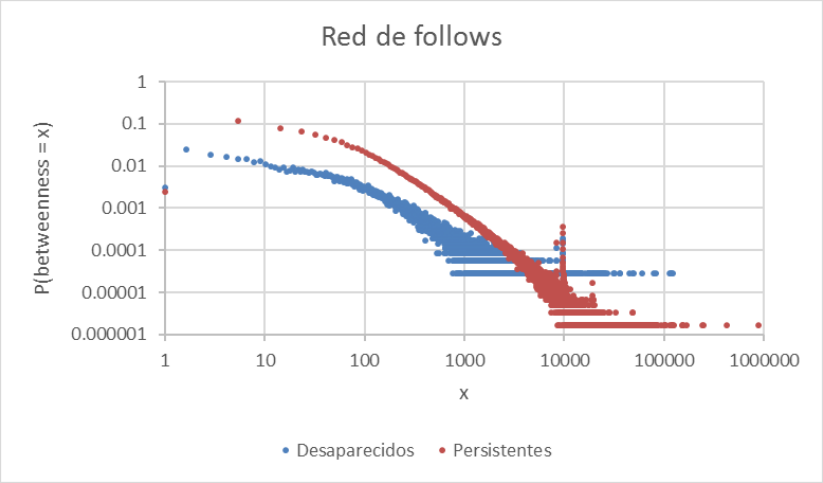

Figura 21. Distribución del betweenness de los enlaces del grafo de follows. Ejes en escala logarítmica.

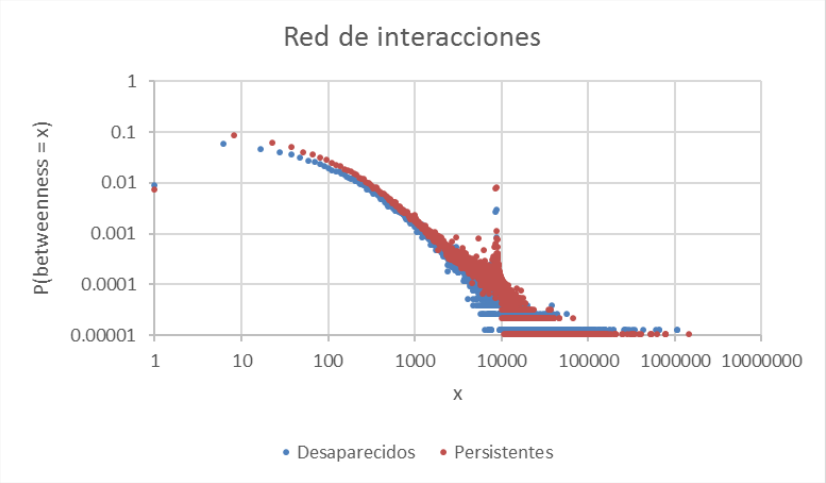

Figura 22. Distribución del betweenness de los enlaces del grafo de follows. Ejes en escala logarítmica.

A.2 Arraigo enlaces

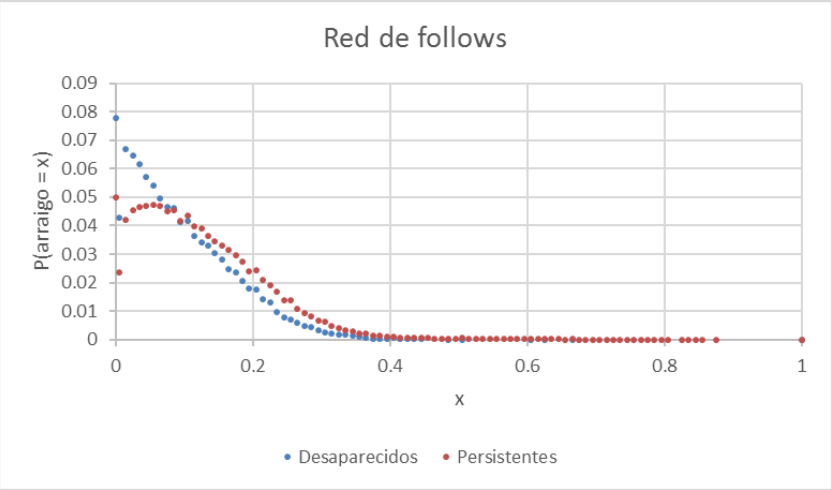

Figura 23. Distribución del arraigo de los enlaces del grafo de follows

Figura 24. Distribución del arraigo de los enlaces del grafo de interacciones

A.3 Reciprocidad enlaces

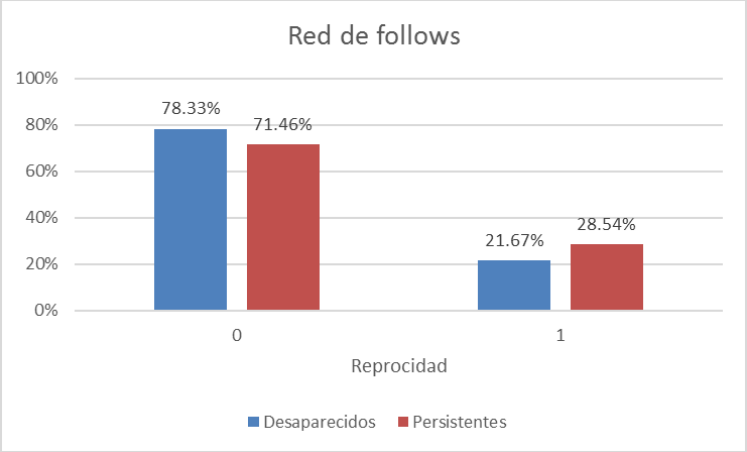

Figura 25. Distribución de la reprocidad de los enlaces del grafo de follows

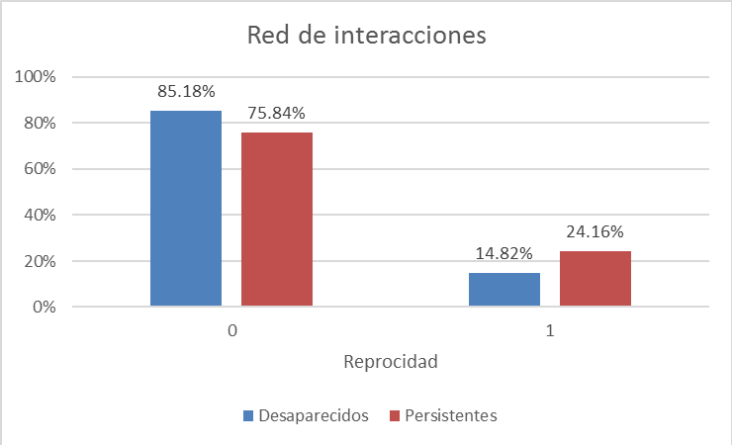

Figura 26. Distribución de la reprocidad de los enlaces del grafo de interacciones

A.4 InDegree nodo origen

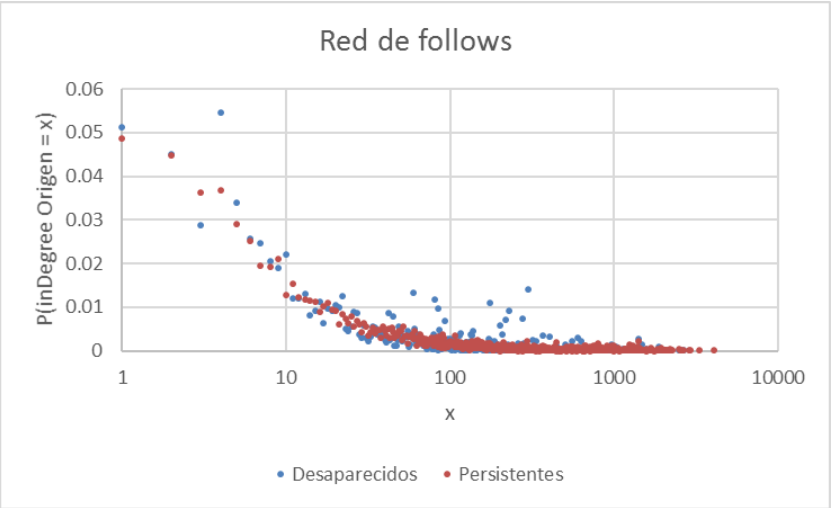

Figura 27. Distribución del indegree del nodo origen de los enlaces del grafo de follows. Eje x en escala logarítmica.

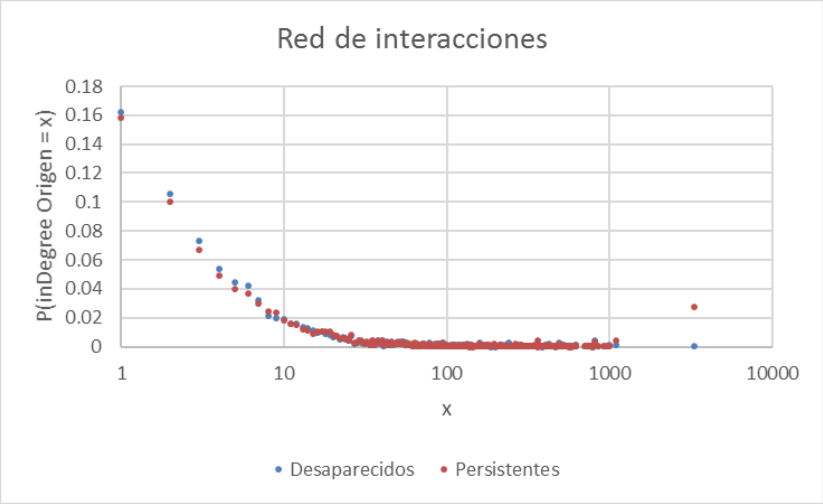

Figura 28. Distribución del indegree del nodo origen de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.5 InDegree nodo destino

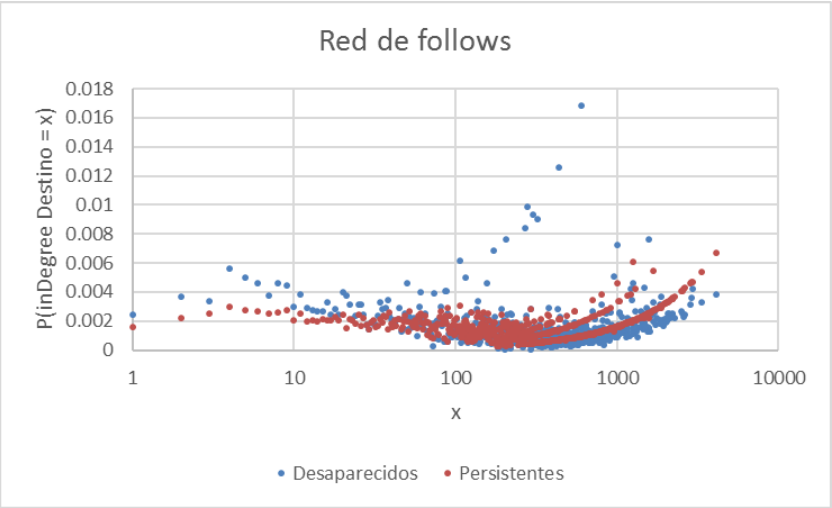

Figura 29. Distribución del indegree del nodo destino de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

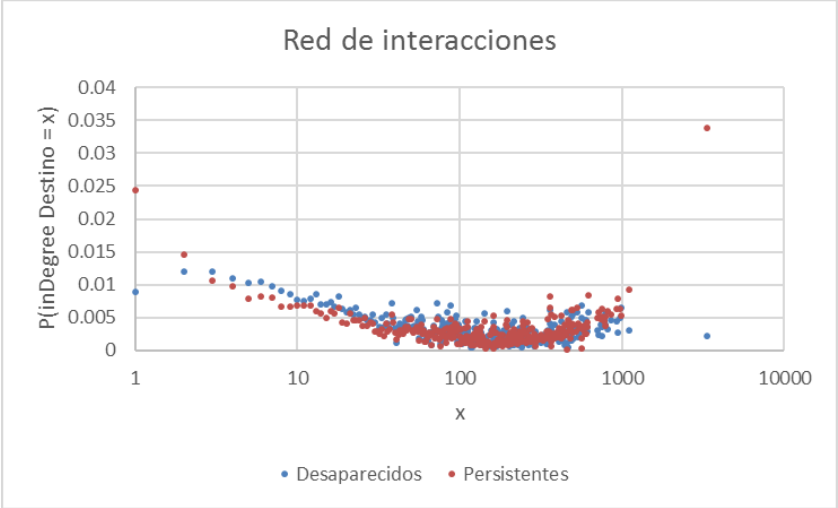

Figura 30. Distribución del indegree del nodo destino los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.6 OutDegree nodo origen

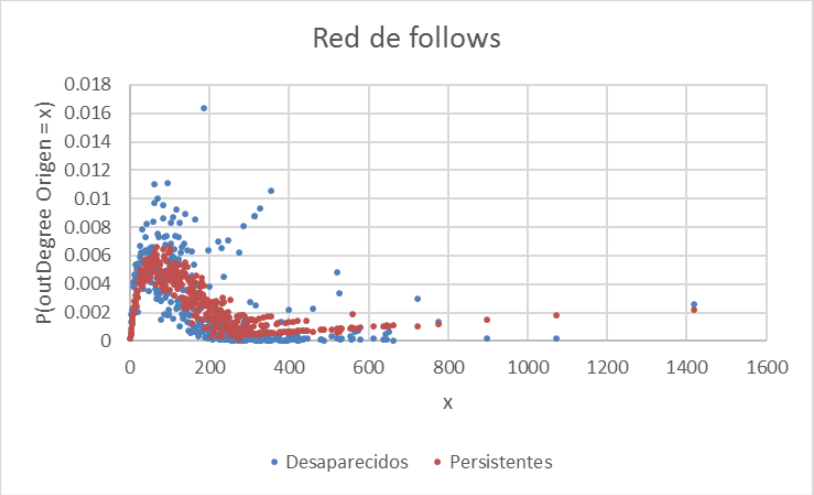

Figura 31. Distribución del outdegree del nodo origen de los enlaces del grafo de follows. Eje x en escala logarítmica.

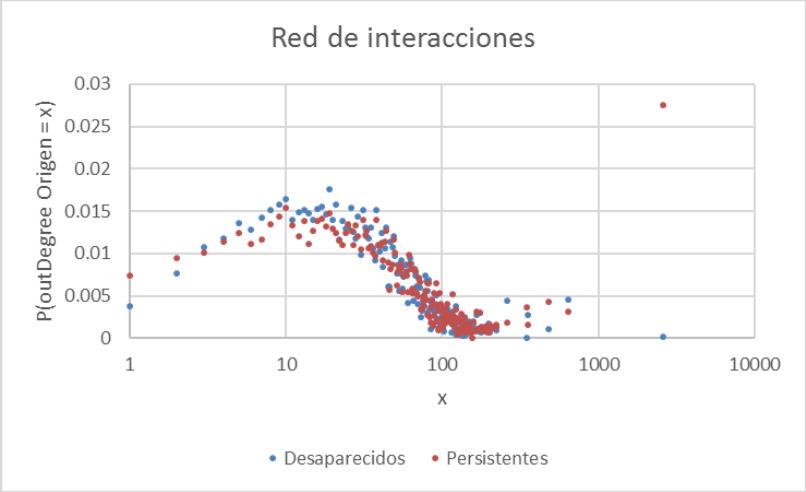

Figura 32. Distribución del outdegree del nodo origen de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.7 OutDegree nodo destino

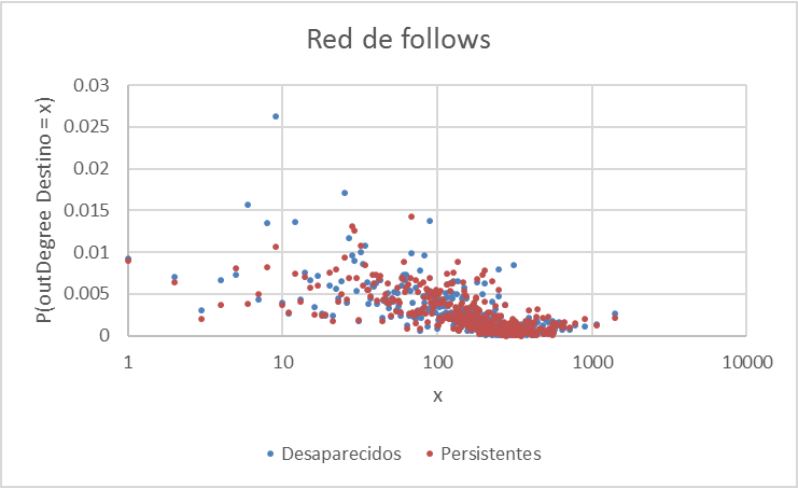

Figura 33. Distribución del outdegree del nodo destino de los enlaces del grafo de follows. Eje x en escala logarítmica.

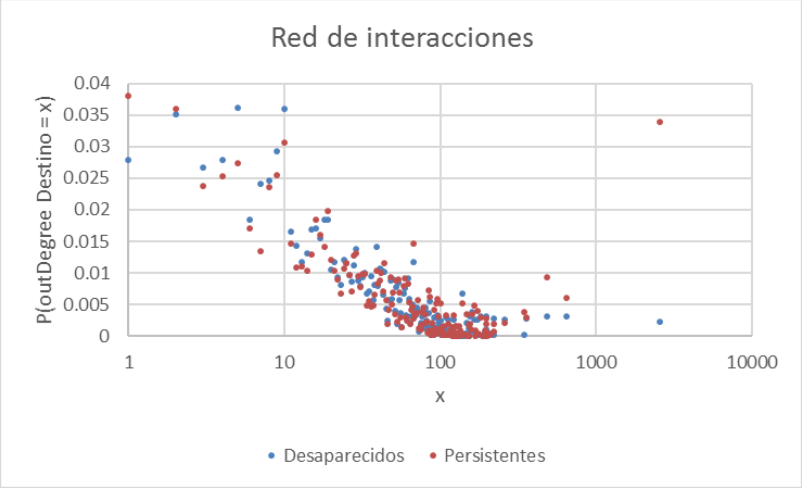

Figura 34. Distribución del outdegree del nodo destino de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.8 Betweenness nodo origen

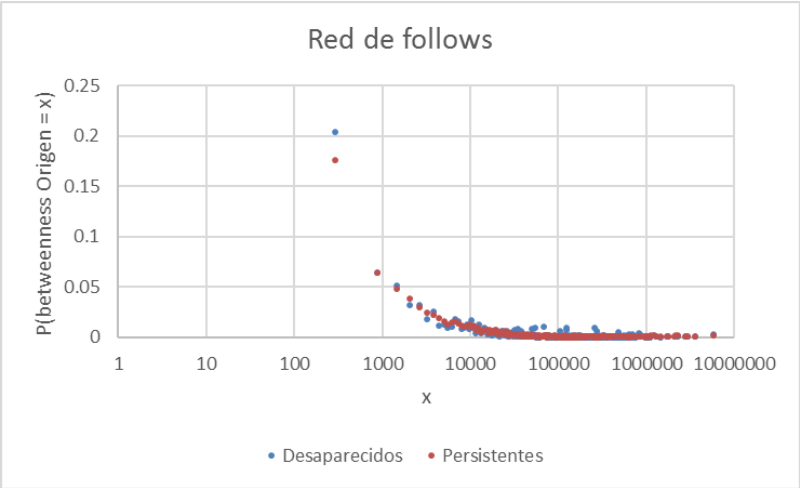

Figura 35. Distribución del betweenness del nodo origen de los enlaces del grafo de follows. Eje x en escala logarítmica.

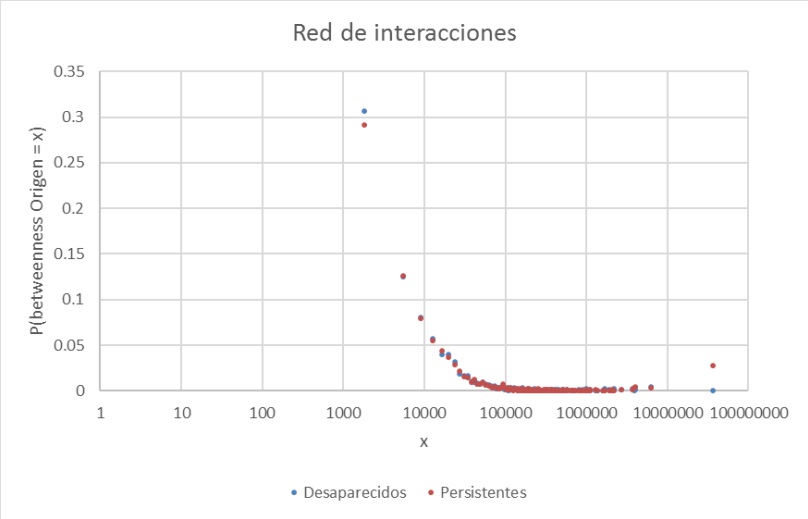

Figura 36. Distribución del betweenness del nodo origen de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.9 Betweenness nodo destino

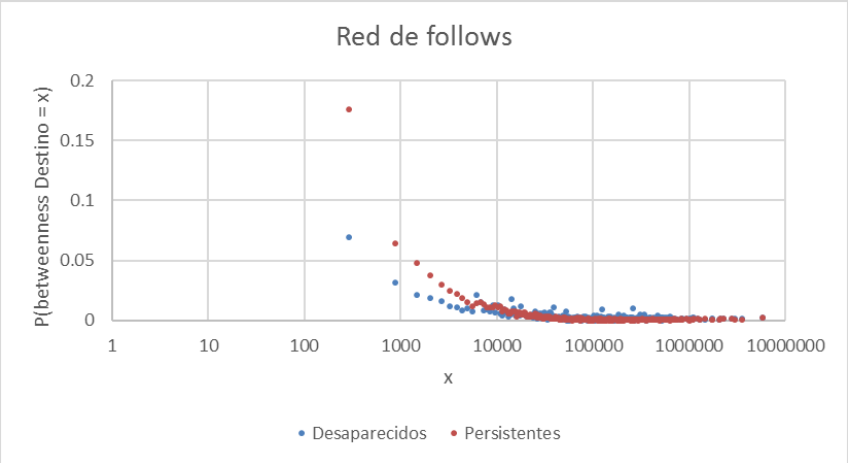

Figura 37. Distribución del betweenness del nodo destino de los enlaces del grafo de follows. Eje x en escala logarítmica.

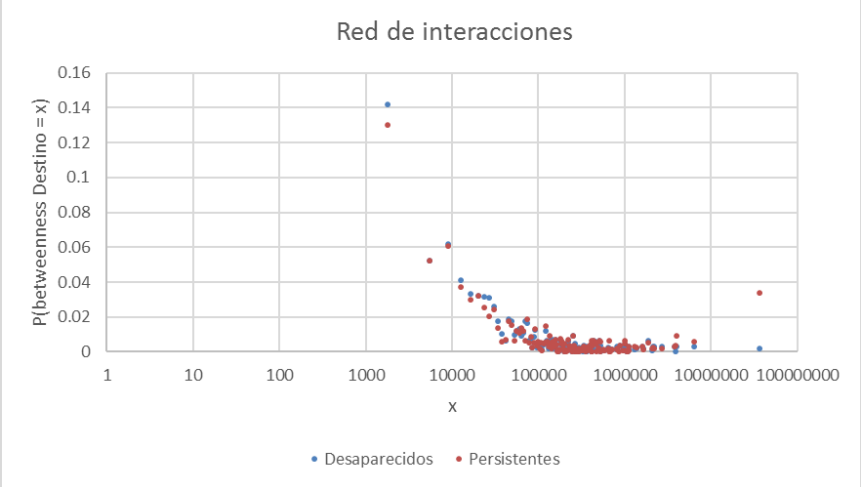

Figura 38. Distribución del betweenness del nodo destino de los enlaces del grafo de interacciones. Eje x en escala logarítmica.

A.10 Coeficiente de clustering nodo origen

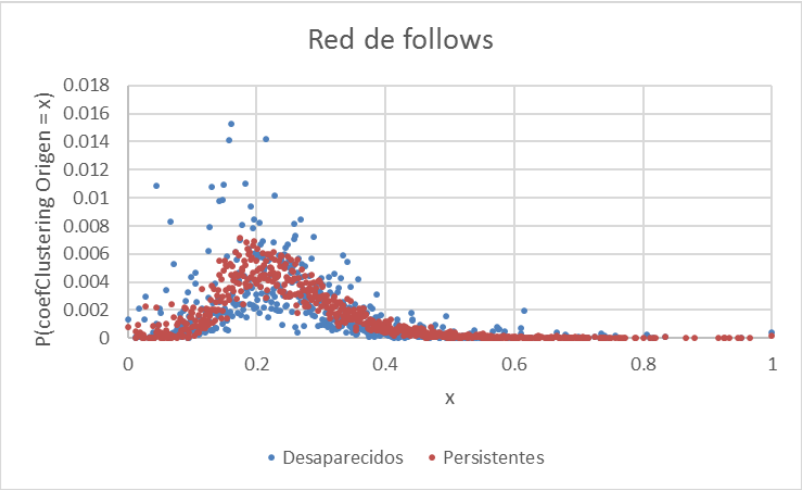

Figura 39. Distribución del coeficiente de clustering del nodo origen de los enlaces del grafo de follows.

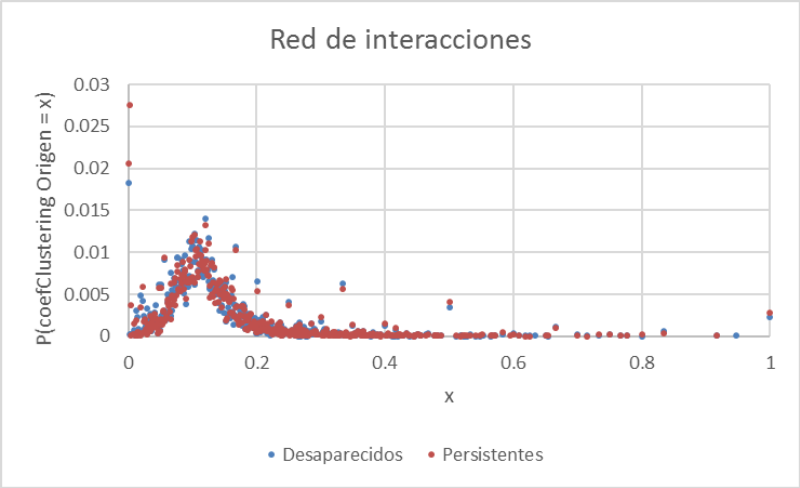

Figura 40. Distribución del coeficiente de clustering del nodo origen de los enlaces del grafo de interacciones.

A.11 Coeficiente de clustering nodo destino

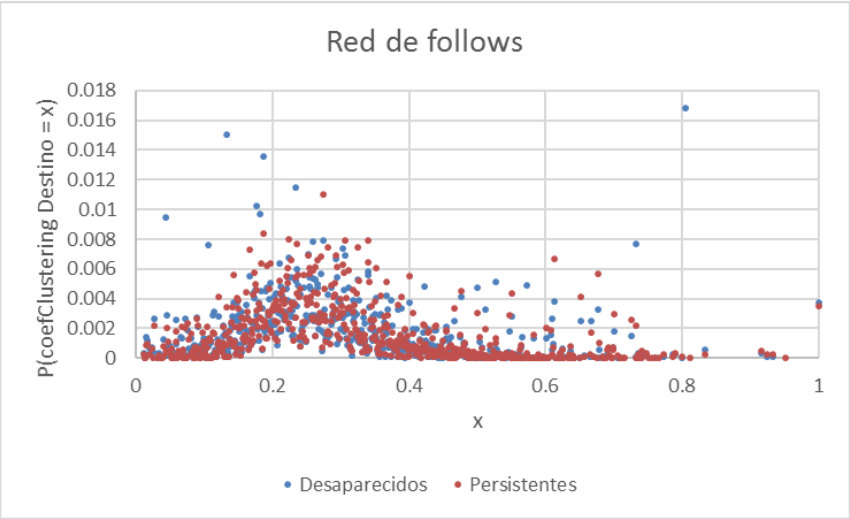

Figura 41. Distribución del coeficiente de clustering del nodo destino de los enlaces del grafo de follows.

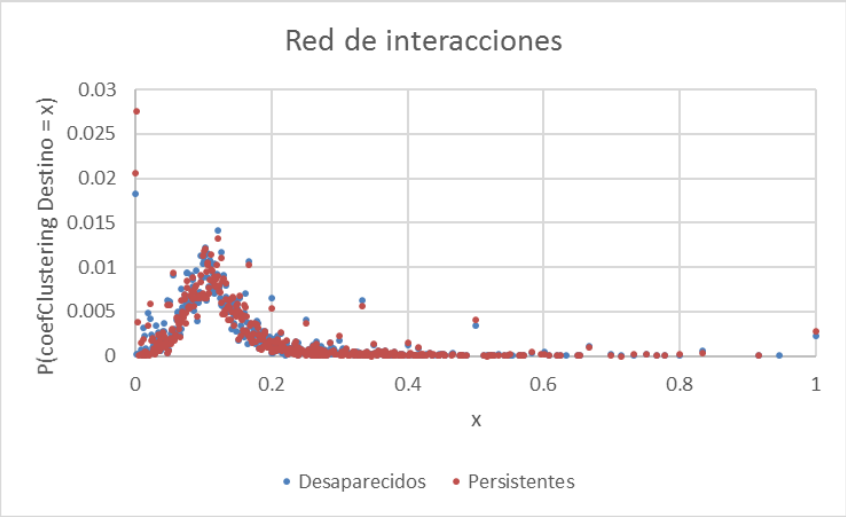

Figura 42. Distribución del coeficiente de clustering del nodo destino de los enlaces del grafo de interacciones.

A.12 PageRank nodo origen

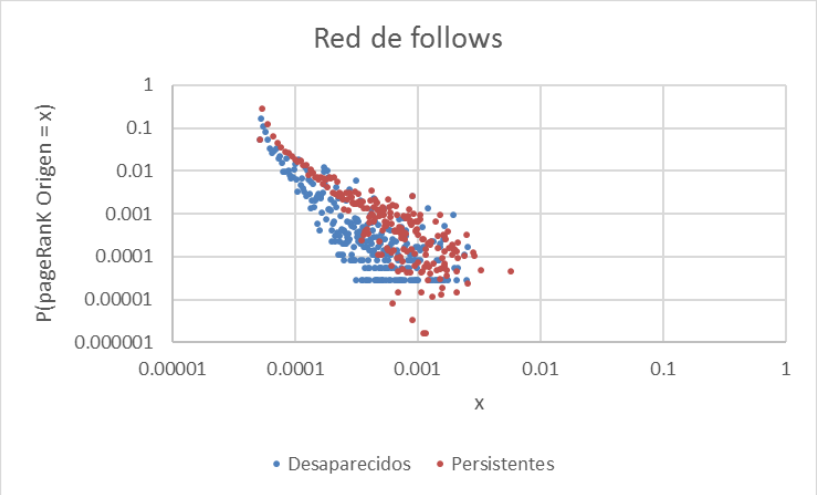

Figura 43. Distribución de PageRank del nodo origen de los enlaces del grafo de follows. Ejes en escala logarítmica.

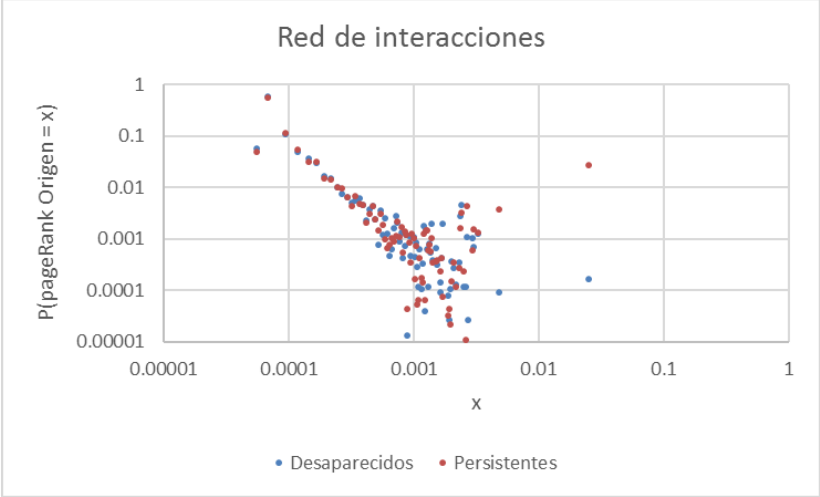

Figura 44. Distribución de PageRank del nodo origen de los enlaces del grafo de interacciones. Ejes en escala logarítmica.

A.13 PageRank nodo destino

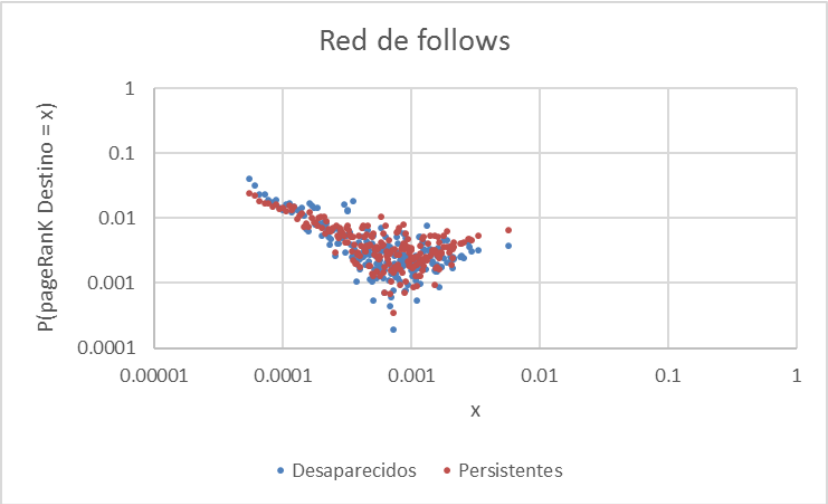

Figura 45. Distribución de PageRank del nodo destino de los enlaces del grafo de follows. Ejes en escala logarítmica.

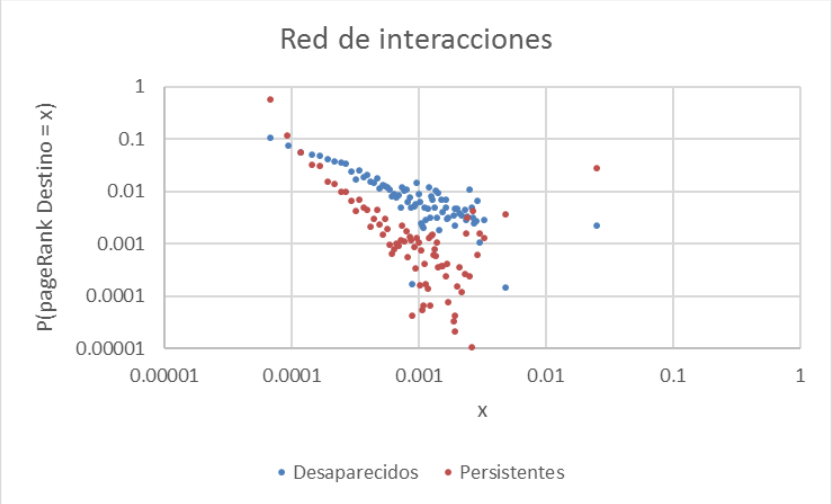

Figura 46. Distribución de PageRank del nodo destino de los enlaces del grafo de interacciones. Ejes en escala logarítmica.

Anexo B. Evaluación de los recomendadores utilizados

B.1 Grafo de follows

Tabla 7. Evaluación de los recomendadores sobre el grafo de follows.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.0664	0.0669	0.0657	0.0623	0.0184	0.0538	0.0843	0.1483
Popularidad	0.0505	0.0541	0.0538	0.0523	0.0120	0.0368	0.0585	0.1060
Popularidad invertido	0.0808	0.0793	0.0774	0.0722	0.0254	0.0697	0.1104	0.1918
Adamic-Adar	0.0424	0.0455	0.0452	0.0460	0.0095	0.0298	0.0477	0.0915
Adamic invertido	0.1174	0.0995	0.0937	0.0834	0.0434	0.0971	0.1423	0.2289
User-based kNN	0.0439	0.0449	0.0448	0.0450	0.0101	0.0297	0.0478	0.0911
User-based kNN invertido	0.0906	0.0871	0.0838	0.0761	0.0285	0.0788	0.1192	0.1982
Vecinos comunes	0.0431	0.0454	0.0458	0.0464	0.0097	0.0299	0.0484	0.0922
Vecinos comunes invertido	0.1196	0.1006	0.0950	0.0837	0.0449	0.0998	0.1473	0.2313
BM25	0.0608	0.0585	0.0566	0.0550	0.0149	0.0405	0.0645	0.1196
BM25 invertido	0.1140	0.0962	0.0899	0.0786	0.0421	0.0912	0.1307	0.2047
Betweenness enlace	0.0753	0.0718	0.0688	0.0660	0.0207	0.0550	0.0854	0.1555
Betweenness enlace invertido	0.0710	0.0655	0.0652	0.0634	0.0221	0.0537	0.0867	0.1559
Betweenness	0.0605	0.0588	0.0569	0.0548	0.0146	0.0406	0.0650	0.1181
Betweenness invertido	0.0869	0.0867	0.0853	0.0770	0.0266	0.0792	0.1262	0.2066
Arraigo	0.0544	0.0537	0.0527	0.0511	0.0156	0.0416	0.0650	0.1183
Arraigo invertido	0.0982	0.0890	0.0836	0.0761	0.0326	0.0784	0.1144	0.1879
Reciprocidad	0.0912	0.0763	0.0710	0.0653	0.0292	0.0685	0.1022	0.1756
Reciprocidad invertido	0.1198	0.0997	0.0923	0.0816	0.0404	0.0881	0.1278	0.2059
InDegree	0.0505	0.0541	0.0538	0.0523	0.0120	0.0368	0.0585	0.1060
InDegree invertido	0.0808	0.0793	0.0774	0.0722	0.0254	0.0697	0.1104	0.1918
OutDegree	0.0633	0.0621	0.0609	0.0589	0.0165	0.0467	0.0751	0.1360
OutDegree invertido	0.0841	0.0814	0.0818	0.0759	0.0245	0.0687	0.1142	0.1948
Coeficiente Clustering	0.1201	0.0890	0.0777	0.0689	0.0404	0.0775	0.1061	0.1720
Coeficiente Clustering invertido	0.0854	0.0787	0.0744	0.0677	0.0260	0.0669	0.1009	0.1706
PageRank	0.0510	0.0549	0.0546	0.0532	0.0119	0.0368	0.0595	0.1081
PageRank invertido	0.0822	0.0803	0.0777	0.0721	0.0268	0.0722	0.1117	0.1928
NaiveBayes	0.1797	0.1709	0.1660	0.1533	0.0353	0.0979	0.1495	0.2483
Logistic	0.1945	0.1793	0.1710	0.1571	0.0422	0.1066	0.1595	0.2630
RandomForest	0.2257	0.2069	0.1927	0.1712	0.0543	0.1286	0.1846	0.2904

B.2 Grafo de interacciones

Tabla 8. Evaluación de los recomendadores sobre el grafo de interacciones.

Recomendador	P@1	P@3	P@5	P@10	R@1	R@3	R@5	R@10
Aleatorio	0.4314	0.3992	0.3720	0.3148	0.1290	0.3004	0.4049	0.5502
Popularidad	0.2863	0.3329	0.3259	0.2913	0.0924	0.2747	0.3823	0.5342
Popularidad invertido	0.4856	0.4387	0.4028	0.3360	0.1504	0.3221	0.4248	0.5661
Adamic-Adar	0.2737	0.3201	0.3156	0.2814	0.0996	0.2713	0.3784	0.5290
Adamic invertido	0.4950	0.4502	0.4128	0.3431	0.1469	0.3249	0.4297	0.5713
User-based kNN	0.2446	0.3035	0.2977	0.2677	0.0852	0.2553	0.3495	0.4865
User-based kNN invertido	0.4938	0.4385	0.3979	0.3284	0.1463	0.3070	0.3995	0.5270
Vecinos comunes	0.2982	0.3260	0.3179	0.2826	0.1086	0.2742	0.3798	0.5298
Vecinos comunes invertido	0.4935	0.4489	0.4116	0.3420	0.1455	0.3236	0.4284	0.5709
BM25	0.4093	0.3813	0.3561	0.3030	0.1290	0.2930	0.3971	0.5422
BM25 invertido	0.4787	0.4273	0.3888	0.3262	0.1395	0.3103	0.4126	0.5567
Betweenness enlace	0.3283	0.3759	0.3642	0.3157	0.0984	0.2909	0.4015	0.5505
Betweenness enlace invertido	0.4381	0.3973	0.3680	0.3116	0.1411	0.3064	0.4071	0.5500
Betweenness	0.2687	0.3424	0.3378	0.2979	0.0893	0.2798	0.3890	0.5394
Betweenness invertido	0.4580	0.4230	0.3901	0.3282	0.1453	0.3145	0.4173	0.5610
Arraigo	0.4253	0.3815	0.3545	0.3006	0.1349	0.2966	0.3989	0.5405
Arraigo invertido	0.4515	0.4182	0.3876	0.3243	0.1336	0.3042	0.4109	0.5555
Reciprocidad	0.3224	0.3525	0.3429	0.3034	0.0987	0.2823	0.3919	0.5437
Reciprocidad invertido	0.4633	0.4256	0.3917	0.3280	0.1466	0.3178	0.4204	0.5613
InDegree	0.2863	0.3329	0.3259	0.2913	0.0924	0.2747	0.3823	0.5342
InDegree invertido	0.4856	0.4387	0.4028	0.3360	0.1504	0.3221	0.4248	0.5661
OutDegree	0.2683	0.3456	0.3387	0.2990	0.0899	0.2819	0.3892	0.5391
OutDegree invertido	0.4553	0.4193	0.3864	0.3270	0.1459	0.3144	0.4160	0.5608
Coeficiente Clustering	0.4704	0.4261	0.3886	0.3232	0.1428	0.3148	0.4152	0.5551
Coeficiente Clustering invertido	0.4057	0.3878	0.3628	0.3084	0.1174	0.2896	0.3973	0.5450
PageRank	0.2734	0.3384	0.3337	0.2946	0.0896	0.2764	0.3858	0.5361
PageRank invertido	0.4833	0.4391	0.4021	0.3345	0.1501	0.3224	0.4245	0.5644
NaiveBayes	0.4673	0.4260	0.3915	0.3298	0.1440	0.3167	0.4189	0.5620
Logistic	0.4736	0.4315	0.3973	0.3313	0.1453	0.3200	0.4228	0.5643
RandomForest	0.7749	0.6549	0.5724	0.4373	0.2272	0.4195	0.5157	0.6342